Do non-native listeners benefit from speech modifications designed to promote intelligibility for native listeners?

The Listening Talker

An interdisciplinary workshop
on natural and synthetic modification of speech
in response to listening conditions

Edinburgh, 2–3 May 2012

Proceedings

Organisers:

Martin Cooke, Simon King, Bastiaan Kleijn & Yannis Stylianou

Technical Committee:

Paavo Alku, Alan Black, Hynek Bořil, Ann Bradlow, Chris Davis, Thierry Dutoit, Maëva Garnier, Agustín Gravano, Joakim Gustafson, John Hansen, Peter Howell, Jean Krause, Gernot Kubin, Philip Loizou, Ewen MacDonald, Thomas Quatieri, Ann Syrdal, Louis Ten Bosch, Tomoki Toda & Werner Verhelst

Local organisation:

Vincent Aubanel, Cassia Valentini-Botinhao & Yan Tang

With the support of:

The LISTA project http://listening-talker.org/

Contents

Torsten Dau

Danish Technical University

Modeling speech intelligibility in adverse conditions

In everyday life, the speech we listen to is often mixed with many other sound sources as well as reverberation. In such situations, people with normal hearing are able to almost effortlessly segregate a single voice out of the background. In contrast, hearing-impaired people have great difficulty understanding speech when more than one person is talking, even when reduced audibility has been fully compensated for by a hearing aid. The reasons for these difficulties are not well understood. This presentation highlights recent concepts of the monaural and binaural signal processing strategies employed by the normal as well as impaired auditory system. Jørgensen and Dau [(2011). J. Acoust. Soc. Am. 130, 1475-1487] proposed the speech-based envelope power spectrum model (sEPSM) in an attempt to overcome the limitations of the classical speech transmission index (STI) and speech intelligibility index (SII) in conditions with nonlinearly processed speech. Instead of considering the reduction of the temporal modulation energy as the intelligibility metric, as assumed in the STI, the sEPSM applies the signal-to-noise ratio in the envelope domain (SNRenv). This metric was shown to be the key for predicting the intelligibility of reverberant speech as well as noisy speech processed by spectral subtraction. However, the sEPSM cannot account for speech subjected to phase jitter, a condition in which the spectral structure of speech is destroyed, while the broadband temporal envelope is kept largely intact. In contrast, the effects of this distortion can be predicted successfully by the spectro-temporal modulation index (STMI) [Elhilali et al., (2003). Speech Commun. 41, 331-348], which assumes an explicit analysis of the spectral modulation energy. However, since the STMI applies the same decision metric as the STI, it fails to account for spectral subtraction. The results from the different modeling approaches suggest that the SNRenv might be a key decision metric while some explicit across-frequency pre-processing seems crucial to extract relevant speech features in some conditions.

Valerie Hazan

Dept. of Speech, Hearing and Phonetic Sciences, UCL, London, UK

The effect of speaker-listener interaction on speech production in adverse listening conditions

Speakers interacting with interlocutors who are having difficulty understanding them due to a hearing impairment or adverse listening conditions need to make adaptations to their speech to maintain effective communication despite the adverse environment. At the same time, according to Lindblom's Hyper-Hypo model of Speech Production (Lindblom, 1990), talkers will tend to keep articulatory effort to the minimum level needed for effective communication. In our recent study, we investigated the adaptations that speakers make in such situations. The LUCID corpus (Baker and Hazan, 2011) includes dialogs produced by 40 speakers of SSBE while resolving a set of 'spot the difference' picture tasks in different communicative conditions. In some, the two talkers carrying out the task could hear each other normally while in others, one talker heard the other via (a) a three-channel noise-excited vocoder (i.e. a simulation of a cochlear implant), (b) multibabble noise or (3) a language barrier (L2 speaker). Crucially, we analysed the speech of the talker in the pair who was hearing normally, but who had to adapt his or her speech in response to the listening difficulties of their interlocutor. Analyses were made of global and segmental acoustic-phonetic measures as well as measures of lexical variety and communication efficiency. The communication barriers elicited perceptually-clearer speech in the talker not directly experiencing the interference, and the adaptations made varied with the type of communication barrier that the interlocutor was experiencing. Correlations in clarity ratings for samples of spontaneous speech produced in the different conditions suggest that talkers' ranking in terms of their inherent clarity persists across speaking styles. However, weak correlations between acoustic-phonetic measures and measures of communication efficiency in adverse conditions suggest that talkers used a range of strategies to clarify their speech. These data provide further evidence that speech production is finely attuned to the needs of the interlocutor and that much is to be gained by analysing speech produced with communicative intent.

Richard Heusdens

Delft University of Technology

Speech intelligibility improvement using a perceptual distortion measure

In this talk we present a speech pre-processing algorithm to improve the speech intelligibility in noise for the near–end listener. The algorithm improves the intelligibility by optimally redistributing the speech energy over time and frequency for a perceptual distortion measure, which is based on a spectro-temporal auditory model.

Perceptual models exploiting auditory masking are frequently used in audio and speech processing applications like coding and watermarking. In most cases, these models only take into account spectral masking in short-time frames. As a consequence, undesired audible artifacts in the temporal domain may be introduced (e.g., pre-echoes). In this talk we discuss a new low-complexity spectro-temporal distortion measure. The model facilitates the computation of analytic expressions for masking thresholds, while advanced spectro-temporal models typically need computationally demanding adaptive procedures to find an estimate of these masking thresholds. We show that the proposed method gives similar masking predictions as an advanced spectro-temporal with only a fraction of its computational power. This auditory model can be used to improve speech intelligibility. Since it takes into account short-time information, transients will receive more amplification compared to stationary vowels, which is beneficial for improving intelligibility in noise. The proposed method is compared to the noisy unprocessed speech and two reference methods by means of an intelligibility listening test. The results show that the proposed method leads to a statistically significant improved speech intelligibility and, at the same time, to improved speech quality compared to the noisy speech.

Distinguished IEEE Lecture

James Johnston

DTS Inc.

Articulation in the presence of noise

This talk will introduce the basic issues in articulation (i.e. understanding speech) in the presence of noise. First, the presentation will provide a quick overview of the psychology, which will point out differences in articulation that can arise due to attention or expectation. Then, a quick mention of auditory masking, combined with auditory filtering, will explain what is necessary to get the peaks in human speech (the formants) above the masking level in a fashion that the auditory system and brain can filter them out of the background. Finally, a discussion of binaural hearing, and how onset helps with disambiguating sounds in the presence of noise, will complete the talk. Due to the broad nature of the particular subject, the talk will be at a higher, more conceptual level, addressing concepts rather than extensive details.

Hideki Kawahara

Wakayama University

A new dimension of voice quality manipulation

Voice quality plays important roles in communication. It provides additional communication channels for non- and paralinguistic information. Extension of a multi-aspect temporally variable morphing framework based on TANDEM-STRAIGHT, by introducing a new texture-related parameters, enables modification of this communication channels while preserving naturalness of the original speech. This talk starts from a brief summary of current state of STRAIGHT and then, introduces finer analysis and synthesis procedures of excitation source related parameters which are basis of this extension. Signal processing aspects such as new instantaneous frequency representation, F0 extraction with higher temporal resolution, parametric representation of aperiodic components and higher statistical aspect such as sound texture provide basis and motivation of this extension.

Roger K. Moore

University of Sheffield

Some insights into talker-listener-environment coupling, energetics and the contrastive particulate structure of spoken language

Back in the early 1980s, my small research group at the Royal Signals and Radar Establishment (consisting of myself, Martin Russell and Mike Tomlinson) were investigating advanced forms of dynamic time warping (DTW). In particular, we were studying the detailed spectro-temporal relationships revealed by DTW between different versions of the same and contrasting utterances, and we came up with two key techniques for modelling the patterning that we observed: 'timescale variability analysis' (TVA) and 'discriminative networks' (DNs). TVA was a forerunner of what became widely known as 'duration modelling', and DNs were effectively a very early form of sub-word modelling. All of the research was conducted using speech that had been parameterised using the front-end of a military-specification channel vocoder (effectively a 27 channel filter bank). The vocoder not only provided the advantage of real-time speech analysis (so we were able to build real-time ASR systems), but it also offered the bonus that any manipulated speech patterns could be replayed through the channel vocoder synthesiser – and that is what we did on a regular basis, not to generate speech per se, but simply to understand the pattern structures that were embedded in our early statistical models. We were therefore quite surprised when we discovered that we could vary the generated output along various continua automatically such that one word could be transformed to sound like another or, much more interestingly, that a word could be transformed to sound less like another, and that the latter manipulations sounded clearer (as if the speaker was making more effort)! We had, of course, stumbled across a practical demonstration of what Bjorn Lindblom would subsequently publish as his theory of hyper and hypo speech (H&H).

Since those early days, I have held a continuing belief that talkers actively manage their speech production to suit the communicative context (including the listener(s) and the environment), and that such teleological behaviour was the source of much unexplained variability. It is for this reason that Lindblom's H&H theory figures strongly in my 'predictive sensorimotor control and emulation' (PRESENCE) model of speech in which I introduce the concept of 'reactive speech synthesis' – a synthesiser that dynamically adjusts its output as a function of the perceived effect on the listener.

In this talk I will discuss my current thinking in this area, touching on Robin Hofe's investigation into H&H using 'AnTon' (his animatronic tongue and vocal tract) and Mauro Nicolao's research into 'speech synthesis by analysis', but also speculating about (i) the wider implications of dynamic coupling between talkers, listeners and their communicative environments, (ii) the fundamental role that energetics plays in conditioning the behaviour of living systems, and (iii) the special consequences for the evolution of a high information-rate low degree-of-freedom system such as spoken language.

Russell, M. J., Moore, R. K., & Tomlinson, M. J. (1983). Some techniques for incorporating local timescale variability information into a dynamic time warping algorithm for automatic speech recognition, IEEE Int. Conf. on Acoustics, Speech and Signal Processing. Boston.

Moore, R. K., Russell, M. J., & Tomlinson, M. J. (1983). The discriminative network; a mechanism for focusing recognition in whole-word pattern matching, IEEE Int. Conf. on Acoustics, Speech and Signal Processing. Boston.

Moore, R. K. (2007). Spoken language processing: piecing together the puzzle. Speech Communication, 49, 418-435.

Moore, R. K. (2007). PRESENCE: A human-inspired architecture for speech-based human- machine interaction. IEEE Trans. Computers, 56(9), 1176-1188.

Hofe, R., & Moore, R. K. (2008). Towards an investigation of speech energetics using 'AnTon': an animatronic model of a human tongue and vocal tract. Connection Science, 20(4), 319–336.

Moore, R. K., & Nicolao, M. (2011). Reactive speech synthesis: actively managing phonetic contrast along an H&H continuum, 17th International Congress of Phonetics Sciences (ICPhS). Hong Kong.

Martin Pickering

University of Edinburgh

An integrated theory of language production and comprehension

Current accounts of language processing treat production and comprehension as quite distinct. I reject this dichotomy. In its place, I propose that producing and understanding are tightly interwoven, and this interweaving underlies people's ability to predict themselves and each other. Based on accounts of action, action perception, and joint action in which action and perception are interwoven to support prediction, I develop analogous accounts of production, comprehension and interactive language. Specifically, I propose that people predict their own utterances at different levels of representation (semantics, syntax, and phonology), and that they covertly imitate and predict their partner's utterances.

Bastian Sauert & Peter Vary

RWTH Aachen University, Germany

Listening enhancement for mobile phones: how to improve intelligibility in a noisy environment

Mobile telephony is often conducted in the presence of strong acoustical background noise such as traffic or babble noise. In this situation, the near-end listener perceives a mixture of the clean far-end (downlink) speech and the acoustical background noise from the near-end and thus experiences an increased listening effort and a possibly reduced speech intelligibility.

While the acoustical background noise signal cannot be influenced, the received clean far-end speech signal can be manipulated by signal processing techniques for reducing the listening effort and for improving the speech intelligibility. We call this approach near-end listening enhancement.

A reasonable objective optimization criterion is to maximize the Speech Intelligibilty Index (SII). The optimization has to take into account constraints arising from the underlying psychoacoustical model of perception and from the limitations of small loudspeakers. The optimization approach and the solutions will be presented.

Alternative time-domain and frequency-domain implementation structures with uniform and non-uniform spectral resolution will be discussed. The experimental setup using a dummy head will be described. Audio examples will be demonstrated.

Furthermore, the applicability in digital hearing aids, car radios, and in-car communication systems will be addressed.

Junichi Yamagishi

The Centre for Speech Technology Research, University of Edinburgh

HMM-based speech synthesis adapted to listeners' and talkers' conditions

It is known that the intelligibility of state-of-the-art hidden Markov model (HMM) generated synthetic speech can be comparable to natural speech in clean environments. However, the situation is quite different if the listener's and/or talker's condition differ. If the environment of the listener is noisy, most often natural speech is still more intelligible than synthetic speech. If the condition of the talker is disordered due to vocal disabilities such as neurological degenerative diseases, the talker's speech may be unintelligible even in clean environments.

In this talk, we introduce our recent approaches to these problems. To improve the intelligibility of synthetic speech in noise, we have proposed two promising approches based on statistical modelling and signal processing. In the former statistical modelling approach, we use speech waveforms and articulatory movements recorded in parallel by electromagnetic articulography and try to create hyper-articulated speech from normal speech by manipulating articulatory movements predicted from HMM [1]. The latter signal processing approach is a new cepstral analysis and transformation method [2] based on an objective intelligibility measure for speech in noise, the Glimpse Proportion measure [3]. This new method aims to modify the spectral envelope of speech in order to increase the intelligibility of speech in noise by modifying the clean speech. Finally we mention other work, in which we create natural and intelligible synthetic voices even from disordered unintelligible speech of individuals suffering from motor neurone disease [4].

[1] Z-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang "Integrating Articulatory Features into HMM-based Parametric Speech Synthesis," IEEE Audio, Speech & Language Processing, vol. 17, no. 6, pp. 1171–1185, August 2009

[2] C. Valentini-Botinhao, R. Maia, J. Yamagishi, S. King, and H. Zen, "Cepstral analysis based on the Glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise," Proc. ICASSP 2012

[3] M.Cooke,"A glimpsing model of speech perception in noise," J. Acoust. Soc. Am., vol. 119, no. 3, pp. 1562–1573, 2006

[4] J. Yamagishi, C. Veaux, S. King and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction," invited review. Acoustical Science & Technology, vol. 33, pp. 1–5, January 2012 http://www.jstage.jst.go.jp/browse/ast/33/1/_contents

A Study on Combined Effects of Reverberation and Increased Vocal Effort on ASR

Hynek Bořil, Seyed Omid Sadjadi, and John H.L. Hansen

Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, U.S.A.
{hynek, sadjadi, john.hansen}@utdallas.edu

Abstract

This study analyzes the individual and combined effect of room reverberation and increased vocal effort on automatic speech recognition. Robustness of several state-of-the-art front-end feature extraction strategies and normalizations to these sources of speech signal variability is evaluated in the context of large and small vocabulary recognition tasks on American English and Czech speech corpora. For the large vocabulary task, speech material from the UT-Scope database comprising American English utterances is used. The Czech speech samples are drawn from the CLSD'05 data corpus and used for the small vocabulary tasks. Both databases contain neutral as well as increased vocal effort recordings. Simulated reverberant test conditions are generated using measured room impulse responses from the AIR database and utilized in the evaluations. It is shown that the robustness of a common MFCC front-end to reverberation and increased vocal effort can be considerably improved when paired with cepstral gain normalization and modified RASTA filtering. A combination of recently proposed mean Hilbert envelope coefficients and modified RASTA is found to provide balanced performance across all reverberation and vocal effort conditions.

1. Introduction

Room reverberation can cause various destructive impacts on spectro-temporal characteristics of speech signals, most notably including temporal smearing, filling dips and gaps in the temporal envelope, increasing the prominence of low-frequency energy, and flattening the formant transitions. These impacts have been categorized as self- and overlap-masking effects [1]. The self-masking effect is caused by early sound reflections in the room that arrive at the receiver (ear or microphone) within 50-80 ms after the direct sound. The overlap-masking effect on the other hand is resulted from late echos (or reflections) which tend to smear the direct sound over time and mask succeeding sounds. It has been shown that the overlap-masking effect of reverberation is the primary cause of degraded speech recognition performance in both human listeners [1] and automatic speech recognizers [2, 3].

In addition to signal distortion, room reverberation may result in increased vocal effort of the speakers [4]. This is due to the fact that room reverberation decreases speech quality and intelligibility, which in turn induces changes in the auditory feedback process. Consequently, speakers increase their vocal effort to compensate for the drop in intelligibility. This increase in vocal effort, which is a function of both reverberation time (aka T_{60}) and talker-to-listener distance [4], has been shown to be a major source of speech signal variability that can ultimately deteriorate performance of ASR.

Hence, in a reverberant environment, an ASR system has to struggle with not only the signal distortions, but also the signal variability due to the increased vocal effort which is induced by reverberation. There have been several research attempts that considered individual impacts of room reverberation [2, 3, 5–7] and increased vocal effort [8, 9] on ASR, and reported compensation strategies to alleviate these impacts. However, to the best of our knowledge, this study is one of the first to consider the individual as well as the combined effects of reverberation and increased vocal effort on ASR. In addition, robustness of various conventional and recently proposed feature extraction/compensation techniques are evaluated in the context of both small and large vocabulary ASR tasks under reverberation, increased vocal effort, and their combination. In particular, motivated by their encouraging performance in speaker identification (SID) under reverberation, the recently proposed mean Hilbert envelope coefficient (MHEC) features [10] are benchmarked against traditional MFCC preceded by long-term log spectral subtraction (LTLSS) [3] and Gammatone subband based non-negative matrix factorization (NMF) [7], as well as MFCC implemented in ETSI advanced front-end (AFE) [11], in our ASR experiments. The feature extraction schemes are paired with a number of popular cepstral normalizations and also recently proposed RASTALP temporal filtering.

*This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067.

2. Mean Hilbert Envelope Coefficients: MHEC

MHEC features have been shown to be an effective alternative to MFCCs for robust SID and ASR tasks under reverberant mismatched conditions [10, 12]. Here, we briefly describe the procedure for MHEC extraction.

First, the pre-emphasized reverberant speech signal is analyzed through a 26-channel Gammatone filterbank. Next, since we are mostly interested in slowly varying amplitude modulations rather than the fine structure, in each channel the Hilbert envelope is calculated and smoothed using a low-pass filter with a cut-off frequency of 20 Hz. In the next stage, the low-pass filtered envelope is blocked into frames of 25 ms duration with a skip rate of 10 ms. To estimate the temporal envelope amplitude in each frame, the sample mean is computed. Note that the sample mean is a measure of the spectral energy at the center frequency of each channel, and therefore overall provides a short-term spectral representation of the speech signal. Next, in each channel, the envelope trajectories are normalized using the long-term average computed over the entire utterance. This stage, which is called subband normalization (SN), functions as an automatic gain control (AGC) and is used to suppress any spectral coloration effect of the reverberation (or the self-masking effect) in different frequency channels. Up to this stage, only the self-masking effect which is due to early reflections has been suppressed. The overlap-masking effect, which is the long-term effect of reverberation and due to late reflections, can be modeled as an uncorrelated additive noise [6], and hence can be compensated via spectral subtraction [13]. The output of this stage represents an estimate of the clean anechoic speech spectrum. In the last stage, natural logarithm is applied to compress the dynamic range of spectral coefficients and followed by the DCT to obtain cepstral features. Here, only the first 13 coefficients (including c_0) are retained after DCT. The final output is a matrix of 13-dimensional cepstral features, entitled the mean Hilbert envelope coefficients (MHEC).

3. Feature Normalizations

Feature normalizations are typically used to transform incoming signal towards characteristics learned by the acoustic models. Depending on the type of normalization, detrimental effects of environmental acoustics, ambient noise, channel variations, as well as variations introduced by speakers can be addressed. In our study, several normalizations that were reported to increase robustness to channel variations and increased vocal effort are evaluated. It is noted that room reverberation can be viewed as a form of a convolutional distortion and as such, can be in part addressed by channel-oriented normalizations. However, in many instances, room impulse responses tend to have long tails compared to typical telephone channel responses – a fact reducing the effectiveness of normalizations operating on the level of short-term win-

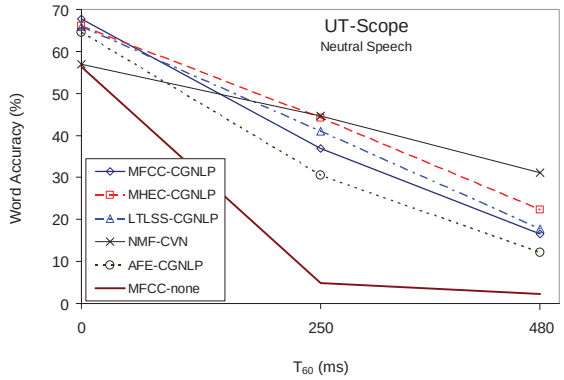


Figure 1: UT-Scope LVCSR; impact of reverberation on neutral speech recognition.

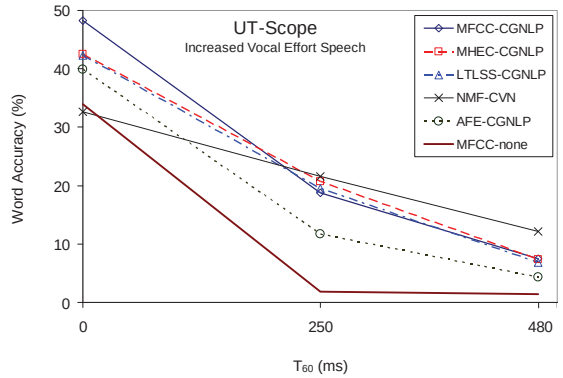


Figure 2: UT-Scope LVCSR; impact of reverberation on increased vocal effort speech recognition.

dows. The following feature normalizations are considered: *Distribution normalizations*: cepstral mean normalization (CMN), cepstral mean/variance normalization (CVN), Gaussianization (feature warping, *warp*) [14], histogram equalization (HEQ) [15], cepstral gain normalization (CGN) [16], and recently established quantile-based cepstral dynamics normalization (QCN) [17]. *Temporal filtering*: Relative spectral (RASTA) filtering [18] and recently proposed modified low-pass RASTA filtering (RASTALP) [9].

In particular, RASTA has been reported to have a potential to reduce the impact of reverberation on ASR [5]. In our previous study, RASTALP – a modified RASTA filter approximating the low-pass component of the original RASTA [9] and the high-pass portion by CMN or other segment-based normalizations [9, 19] was presented. Compared to the original high order RASTA filter, RASTALP requires significantly lower (2^{nd}) filter order, which results in considerable reduction of the transient effects typical for RASTA filtering. The combination of CMN–RASTALP outperformed RASTA in LVCSR on neutral and Lombard speech tasks in clean and noisy conditions [19].

4. Experimental Results

Two different speech corpora are utilized in this study – UT-Scope [20] and CLSD'05 [21]. Both databases contain neutral and increased vocal effort speech. The increased vocal effort was induced by exposing the subjects to background noise, yielding a so called Lombard effect speech [22]. Lombard effect results in the increase of vocal effort and mean fundamental frequency, and affects also a number of other speech parameters [8, 21, 23, 24]. While the casue of the vocal effort increase is different for Lombard speech and speech produced in distant speaker-to-listener conditions, in both cases, the speech modifications result from the alteration of the auditory feedback. Due to the physiological mechanisms, the increased vocal effort goes hand in hand with changes of inherently related speech production parameters. Subglottal pressure and tension in the laryngeal musculature in higher vocal effort cause increase of mean fundamental frequency F_0 [25], which has been observed for altered auditory feedback both due to noise (Lombard effect) [8] and distant speaker-to-listener communication [4]. Increased vocal intensity is accompanied by the jaw lowering, which in turn causes an upward shift of the first formant F_1 [26]. Both migration of spectral energy and spectral center of gravity to higher frequencies [23], as well as flattening of the spectral tilt, are also typical for increased vocal effort in loud and Lombard speech [8]. Considering these similarities, clean Lombard speech seems to be a good approximation of the increased vocal effort speech observed in reverberation, and, hence, is used in this study.

4.1. UT-Scope Speech Corpus

The Lombard effect portion of the UT-Scope speech database contains neutral (modal) speech and speech produced with various levels of increased vocal effort [20]. The increased vocal effort was induced by playing three types of noises for subjects through headphones, and speech was captured by a close-talk microphone, yielding high signal-to-noise ratio (SNR) recordings. This allows for analysis of increased vocal effort speech with the inducing noise being excluded from the signal. The noise types used are: (i) highway car noise (speed 65 mph, windows half open) (ii) crowd noise, and (iii) pink noise. Highway and crowd noises were played at 70, 80, and 90 dB sound pressure level (SPL), pink noise at 65, 75, and 85 dB SPL. Sessions from 31 native speakers of American English (25 females, 6 males) are employed in the ASR experiments.

4.2. CLSD'05 Speech Corpus

The Czech Lombard Speech Database (CLSD'05) [21] comprises recordings of neutral speech and speech uttered in simulated noisy conditions (90 dB SPL of car noise). Similar as in UT-Scope, the noise samples were played through headphones, and speech was collected by a close-talk microphone. Sessions from 26 native speakers of Czech (12 females, 14 males) are utilized in the ASR experiments.

4.3. AIR Database

Two different reverberant test conditions are simulated by convolving the speech material with measured room impulse responses (RIR) from the Aachen Impulse Response (AIR) Database [27]. The RIR samples for a meeting room as well as an office are used with dimensions of $8.0 \times 5.0 \times 3.1$ m^3 and $5.0 \times 6.4 \times 2.9$ m^3, and reverberation times of T_{60}=250 ms and T_{60}=480 ms, respectively. Source-to-microphone distance is 2.8 m in the meeting room, and 3.0 m in the office.

4.4. UT-Scope LVCSR Experiments

Both UT-Scope and CLSD'05 ASR systems utilize HTK for acoustic modeling, and the acoustic front-end features contain 13 static cepstral coefficients, including c_0, and their first and second order time derivatives.

A triphone recognizer with an SRILM trigram language model

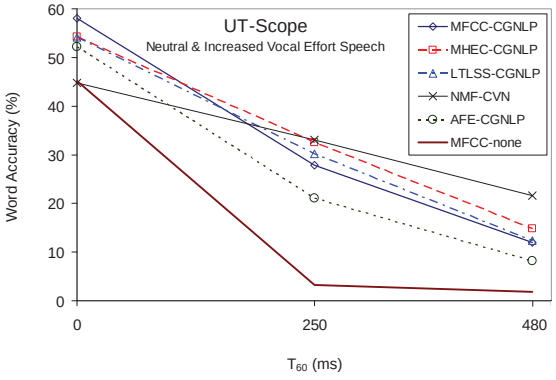


Figure 3: UT-Scope LVCSR; impact of reverberation on speech recognition on pooled neutral and increased vocal effort speech.

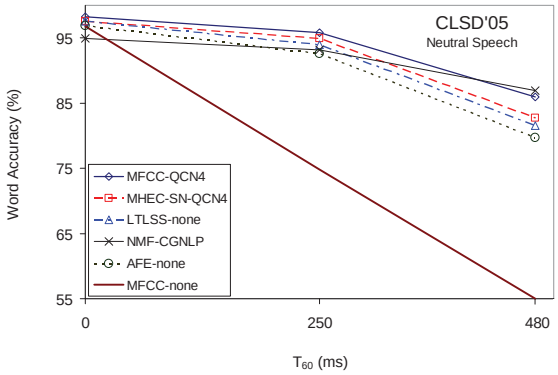


Figure 4: CLSD'05 Digit Recognition; impact of reverberation on neutral speech recognition.

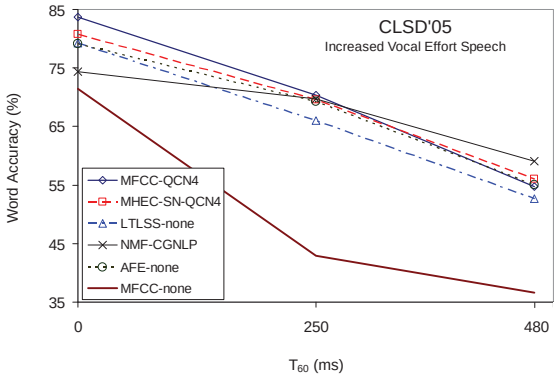


Figure 5: CLSD'05 Digit Recognition; impact of reverberation on increased vocal effort speech recognition.

(LM) is trained on the TIMIT database. Here, 32-mixture TIMIT acoustic models are adapted towards UT-Scope with a combination of maximum likelihood linear regression (MLLR) adaptation, and maximum a posteriori (MAP) adaptation. The adaptation data is drawn from the UT-Scope *clean neutral samples*. Sessions of the adaptation set subjects are withdrawn from the evaluations. The test set contains sessions from 3 male and 19 female subjects. A total of 100 phonetically balanced TIMIT-like sentences produced in the neutral condition, and 20 TIMIT sentences produced in nine noise type/level conditions are available for each subject.

The ASR setups are evaluated on (i) *anechoic sets* – neutral speech and anechoic increased vocal effort speech produced in 70, 80, and 90 dB SPL of simulated highway and crowd noise, and 65, 75, and 85 dB of pink noise (the noise is not present in the recordings); (ii) the previous sets, reverberated with the RIR sample (T_{60} = 250 ms) from the AIR database; (iii) sets from *i* reverberated with the RIR sample (T_{60} = 480 ms) from the AIR database. This totals in 30 evaluation sets. The initial ASR system with MFCC–CVN front-end provides performance of 91.7 % word accuracy (*Acc*) on the anechoic neutral set. Since the focus of this study is on the effects of increased vocal effort and reverberation on acoustic modeling in ASR, the remainder of the paper reports word accuracies with LM being bypassed.

4.5. CLSD'05 Small Vocabulary Experiments

A monophone recognizer is trained on the Czech SPEECON database [28]. The recognizer comprises 43 context-independent monophone models and two silence models. The models are trained on large vocabulary material from the Czech SPEECON database [28]. The task is to recognize 10 Czech digits (16 pronunciation variants) presented in connected digits utterances. The neutral test set comprises a total of 6353 words and the increased vocal effort test set 11663 words. Similar as in the UT-Scope case, the neutral and increased vocal effort sets are presented in the anechoic (i.e., original) and reverberant (T_{60} = 250 ms and T_{60} = 480 ms) conditions.

4.6. Results and Discussion

This section presents the observations made in the UT-Scope LVCSR and CLSD'05 digit recognition experiments. Since the number of the ASR evaluation tasks, as well as the number of feature extraction strategies considered is extensive, in the following paragraphs we attempt to only summarize the overall trends and main outcomes of the experiments.

UT-Scope LVCSR Experiments: In the first step, efficiency of the normalizations from Sec. 3 included in an MFCC front-end was studied for anechoic and reverberated sets comprising pooled neutral and increased vocal effort samples. With increasing reverberation time T_{60}, the ASR performance severely deteriorates for all front-end setups. In all conditions, the combination of CMN and RASTALP (CMN-RASTA) outperformed traditional RASTA. The combinations CGN-RASTALP and QCN4-RASTALP consistently ranked among the top four normalizations in all scenarios, and ten out of twelve best performing front-ends utilized RASTALP filtering. Since CGN-RASTALP preceded, in the terms of word accuracy, QCN4-RASTALP in two out of three scenarios, it is considered to be the most efficient normalization in this evaluation. Detailed results of this experiment can be found in [12].

In the next step, selected feature extraction strategies were evaluated in combination with four normalizations (no normalization, CMN, CVN, and the best performing normalization identified in the previous paragraph – CGN-RASTALP). CMN and CVN are chosen to represent the common choice in many ASR engines. Front-ends mentioned in Sec. 1 and MHEC incorporating spectral subtraction (*MHEC-SS*), sub-band normalization (*MHEC-SN*), or both (*MHEC-SS-SN*) were combined with normalizations and evaluated on anechoic and reverberated sets. On *anechoic data sets*, once combined with any normalization, MFCC reached a superior performance. LTLSS, MHEC, and MHEC-SN ranked second behind MFCC. NMF provided inferior performance to all other front-ends. For *reverberated data* ($T_{60} = 250$ ms), MHEC-SS and MHEC-SN performed best, followed by NMF and MHEC. ETSI-AFE performed inferior to other front-ends. On *reverberated data* ($T_{60} = 480$ ms), NMF established highest accuracy, followed by the four MHEC configurations. CGN_{LP} is most beneficial for all extraction strategies, except for NMF, which benefits most from CVN. Hence, NMF is paired with CVN and all other front-ends employ CGN_{LP} in the subsequent analysis.

Third, feature extraction strategies are paired with their respective 'optimal' normalizations and evaluated separately for neutral, increased vocal effort, and pooled sets in anechoic, T_{60}=250 ms, and T_{60}=480 ms reverberation conditions. For comparison, performance of a baseline MFCC front-end without any normalization (denoted MFCC-none) is also evaluated. As can be seen in Fig. 1 and 2, MFCC-CGN_{LP} establishes the best performance in anechoic conditions for both neutral and increased vocal effort speech while ranking fourth behind NMF, MHEC, LTLSS in $T_{60} = 250$ ms. On the other hand, NMF provides inferior performance in anechoic conditions (even dropping below the baseline performance for increased vocal effort speech) but matches the top-performing front-ends in $T_{60} = 480$ ms and clearly

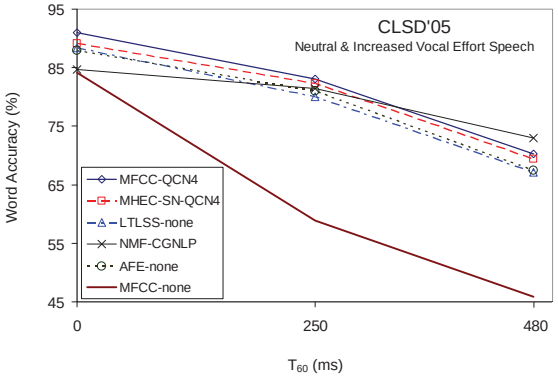


Figure 6: CLSD'05 Digit Recognition; impact of reverberation on speech recognition on pooled neutral and increased vocal effort speech.

dominates in $T_{60} = 480$ ms. Fig. 3 suggests that the MHEC front-end would be the best choice for a recognizer operating in varying reverberation and vocal effort conditions, as MHEC-CGN$_{LP}$ provides the most balanced performance for pooled neutral and increased vocal effort speech in anechoic and reverberated conditions.

CLSD'05 Small Vocabulary Experiments: Unlike in the case of UT-Scope, CLSD'05 experiments focus on the small vocabulary digit recognition task. In addition, CLSD'05 captures Czech spoken language and the recognizer utilizes monophone acoustic models. These differences allow for analysis of how transferable are the observations made in the previous paragraphs to another language and recognition task domain[1]. In the CLSD'05 experiments, the front-end feature extraction strategies (MFCC, MHEC, LTLSS, NMF, AFE) were combined with all normalizations from Sec. 3 but feature warping and histogram equalization (those two provided a suboptimal performance in the initial experiments).

Figures 4, 5, and 6 depict the performance of each feature extraction strategy paired with the respective best performing normalization. It can be seen that the overall ASR performance is considerably higher here due to the simplicity of the task (digit recognition vs. LVCSR). Also, the recognition deterioration is much milder when switching from anechoic to T_{60}=250 ms conditions. Similarly, the small vocabulary recognition accuracy reduces less when switching from neutral to increased vocal effort speech. It can be assumed that the word models in general small vocabulary task are more easily distinguishable in the acoustic feature space and are less affected by the speech deterioration due to reverberation compared to the LVCSR triphone acoustic models.

Surprisingly, all feature extraction strategies in the CLSD'05 task paired with different 'optimal' normalizations than in the UT-Scope task. Similar to UT-Scope, MFCC maintains the best performance on anechoic neutral and increased vocal effort sets. This transfers also to the T_{60}=250 ms condition here. Similar to UT-Scope, NMF dominates in $T_{60} = 480$ ms for both types of speech. On CLSD'05, the combination of MFCC and QCN represents the best choice across most of the conditions, MHEC-SN and QCN being the second best front-end.

It can be seen that while plain MFCC front-end does not deal well with either reverberation or increased vocal effort in both UT-Scope and CLSD'05, it becomes quite competitive when paired with a well chosen normalization – see Fig. 2 in the UT-Scope task and Fig. 4–6 in the CLSD'05 task.

5. Conclusion

This study analyzed the individual as well as combined impacts of reverberation and increased vocal effort on large and small vocabulary recognition in English and Czech languages, respectively. Robustness of several standard and state-of-the-art feature extraction techniques and normalizations was evaluated in the varying reverberation/vocal effort conditions. Several similar trends were observed for both the English large and Czech small vocabulary tasks. In particular, it was observed that ASR performance deteriorated with increasing vocal effort and reverberation time. However, this deterioration is milder in the low vocabulary task, presumably due to the lower confusability of the small vocabulary models in the acoustic feature space. When combined with an 'optimal' normalization, MFCC front-end always outperformed other schemes in anechoic conditions, both for neutral and increased vocal effort speech. In addition, its performance remained competitive with other front-ends in T_{60}=250 ms for increased vocal effort speech in the large vocabulary task as well as both neutral and increased vocal effort speech in the small vocabulary task. NMF front-end was consistently inferior in anechoic conditions, but became competitive in T_{60}=250 ms and dominated in $T_{60} = 480$ ms in all evaluations. Recently established MHEC feature extraction front-end provided well balanced performance in both LVCSR and small vocabulary tasks and state-of-the-art QCN and CGN-RASTALP normalizations ranked among the top choices for most front-ends considered in this study.

[1] In an ideal world, the previous best front-end configuration might be expected to provide a sustained superior performance also here.

6. REFERENCES

[1] A. K. Nabelek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.*, vol. 86, pp. 1259–1265, Oct. 1989.

[2] Q. Lin, C. Che, D.-S. Yuk, L. Jin, B. deVries, J. Pearson, and J. Flanagan, "Robust distant-talking speech recognition," in *Proc. IEEE ICASSP*, vol. 1, May 1996, pp. 21–24.

[3] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition," in *Proc. ICSLP*, Sept. 2002, pp. 2185–2188.

[4] D. Pelegrín-García, B. Smits, J. Brunskog, and C.-H. Jeong, "Vocal effort with changing talker-to-listener distance in different acoustic environments," *J. Acoust. Soc. Am.*, vol. 129, no. 4, pp. 1981–1990, Apr. 2011.

[5] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proc. IEEE ICASSP*, vol. 2, Apr. 1997, pp. 1259–1262.

[6] K. Lebart, J. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica*, vol. 87, pp. 359–366, 2001.

[7] K. Kumar, R. S. B. Raj, and R. M. Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR," in *Proc. IEEE ICASSP*, May 2011, pp. 5448–5451.

[8] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, vol. 20, no. 1-2, pp. 151–173, Nov. 1996.

[9] H. Bořil and J. H. L. Hansen, "UT-Scope: Towards LVCSR under Lombard effect induced by varying types and levels of noisy background," in *Proc. IEEE ICASSP*, Prague, Czech, May 2011, pp. 4472–4475.

[10] S. O. Sadjadi and J. H. L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. IEEE ICASSP*, May 2011, pp. 5448–5451.

[11] "Speech processing, transmission and quality aspects (stq), distributed speech recognition, advanced front-end feature extraction algorithm, compression algorithm," in *ETSI standard document-ETSI ES 202 050 v1.1.1*, 2002.

[12] O. Sadjadi, H. Bořil, and J. H. L. Hansen, "A comparison of front-end compensation strategies for robust LVCSR under room reverberation and increased vocal effort," to appear in *Proc. IEEE ICASSP*, Mar. 2012.

[13] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. ASLP*, vol. 14, pp. 774–784, May 2006.

[14] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. A Speaker Odyssey - The Speaker Recognition Workshop*, Jun. 2001, pp. 213–218.

[15] S. Dharanipragada and M. Padmanabha, "A nonlinear unsupervised adaptation technique for speech recognition," in *Proc. ICSLP*, Oct. 2000, pp. 556–559.

[16] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, "Cepstral gain normalization for noise robust speech recognition," in *Proc. IEEE ICASSP*, vol. 1, May 2004, pp. 209–212.

[17] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Trans. ASLP*, vol. 18, no. 6, pp. 1379–1393, Aug. 2010.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. SAP*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[19] H. Bořil, F. Grézl, and J. H. L. Hansen, "Front-end compensation methods for LVCSR under Lombard effect," in *Proc. INTERSPEECH*, 2011.

[20] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. ASLP*, vol. 17, no. 2, pp. 366–378, Feb. 2009.

[21] H. Bořil, "Robust speech recognition: Analysis and equalization of Lombard effect in Czech corpora," Ph.D. dissertation, CTU in Prague, Czech Rep., http://www.utdallas.edu/∼hynek, 2008.

[22] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, Jan. 1993.

[23] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble and stationary noise," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3261–3275, Nov. 2008.

[24] M. Garnier, "Communication in noisy environments: From adaptation to vocal straining," Ph.D. dissertation, Univ. of Paris VI, France, 2007.

[25] R. Schulman, "Dynamic and perceptual constraints of loud speech," *J. Acoust. Soc. Am.*, vol. 78, no. S1, pp. S37–S37, 1985.

[26] ——, "Articulatory dynamics of loud and normal speech," *J. Acoust. Soc. Am.*, vol. 85, no. 1, pp. 295–312, Jan. 1989.

[27] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. IEEE DSP*, Jul. 2009, pp. 1–5.

[28] D. Iskra, B. Grosskopf, K. Marasek, H. van den Huevel, F. Diehl, and A. Kiessling, "SPEECON – Speech databases for consumer devices: Database specification and validation," in *Proc. of LREC'2002*, 2002, pp. 329–333.

Evaluating the effects of hearing protection on speech production in noisy environments

Douglas S. Brungart, Mary T. Cord, Nancy P. Solomon, Katie Dietrich- Burns, Kim Block

[1]Audiology and Speech Center, Walter Reed National Military Medical Center, Bethesda, MD 02123

douglas.brungart@us.army.mil

Abstract

Many factors can influence the voice levels that talkers choose to use when they initiate conversations in noisy environments, including the desire to be understood clearly, the desire to communicate privately, and the distorted perception of the background noise and of their own voices that can occur when hearing protection devices are worn. In this study, we examined the impact that two different kinds of earplugs and four levels of room noise had on the voice levels of talkers who were asked to privately communicate short verbal messages to a nearby acoustic manikin. Frozen noise samples were used for the room noise, which allowed the speech samples recorded by the manikin to be extracted from the noise by direct waveform subtraction. The results show that talkers wearing earplugs consistently use lower voice levels in noise than they do when not wearing earplugs, even when the earplugs produce relatively little attenuation. This highlights the important impact that earplug-related distortions in the perception of ones own voice (which are commonly referred to as the occlusion effect) can have on the effectiveness of speech communication in noise.

Index Terms: shouted speech, hearing protection, occlusion effect

1. Introduction

Many variables can influence the vocal effort levels that talkers choose to deploy when they are communicating in noisy environments. Usually, the primary goal is to speak loudly enough to be heard by the person to whom they are speaking. However, in most cases, the talker does not want to speak any louder than necessary, both because it requires additional effort to speak loudly and because the talker may be concerned that other people in the environment might overhear the conversation and/or be annoyed by it. In military environments, soldiers may also be concerned about being detected by an adversary if they shout too loudly. Thus, in most cases, talkers in noisy environments are motivated by an underlying goal to speak just loudly enough to be clearly understood by their target audience.

The actual selection of this appropriate voice level is not trivial, even in the best circumstances, because it requires the talker to accurately estimate both the loudness of his or her own voice at the location of the listener and the level of noise at that location. However, this selection process is complicated even further in cases where the person talking is required to use some type of hearing protection. Hearing protection has two important effects on how the environment is perceived by the talker. The first and most obvious impact is that it attenuates the perceived level of noise in the room. However, it also has the more subtle effect of distorting the way the talker perceives his or her own voice. Almost everyone who has used earplugs knows that they make ones own voice sound "boomy" or "hollow", similar to what you might expect to perceive if you were talking with your head inside a barrel. This phenomenon, which is known as the "occlusion effect", is caused by the transmission of the talker's voice through bone conduction into the constricted ear canal space confined by the earplug. Acoustically, its effects can be estimated by placing a microphone in the ear canal and comparing the speech levels measured for a talker speaking at the same level with or without the earplug inserted. Measurements of this type show that the insertion of an earplug can increase the loudness of a talkers voice at low frequencies (250-500 Hz) by 20 dB or more [1]. Occlusion can also be measured subjectively, by obtaining the talkers rating of the hollowness of his or her own voice. Alternatively, bone conduction thresholds can be used to determine the minimal detectable intensity of a 500 Hz tone generated by a bone-conduction transducer with and without the earplug inserted (in which case the insertion of an earplug will produce a 10-20 dB decrease in the minimum detectable threshold).

From these results, one might expect the use of hearing protection to have two consequences for speech production in noise: 1) to reduce the perceived level of noise in the surrounding environment; and 2) to increase the talkers self-perceived vocal loudness for a given amount of vocal effort. Not surprisingly, the net effect of these distortions is a general tendency for talkers using hearing protection in noise to speak more quietly then they normally would if they were speaking in the same environment with open ears. This was demonstrated by Tufts and Frank [2], who asked talkers wearing foam or flange earplugs to produce speech in various levels of simulated noise. However, in order to obtain clean recordings of the talker's voice that were not contaminated by the room noise, that study was conducted in a quiet room with noise-producing headphones worn over the earplugs, rather than in an actual noise-filled room. This made it impossible to evaluate the impact of earmuffs on speech production in noise, and it meant that the "open ear" condition was not truly an "open ear in noise" condition but rather a "noisy headphone in quiet" condition. In this study, we conducted a similar experiment using reproducible frozen noise to adjust the level of a diffuse noise field in an audio test booth. This allowed us to use waveform subtraction to recover the noise spectrum of the talker's voice even in cases where the subject was truly speaking with unoccluded ears in a noisy environment.

2. Hearing Protection Devices

The two earplug conditions tested in the experiment were the two possible configurations of the EAR Combat Arms Earplug (CAE). The CAE is a passive nonlinear earplug system that is designed to protect listeners from high-level impulse noises

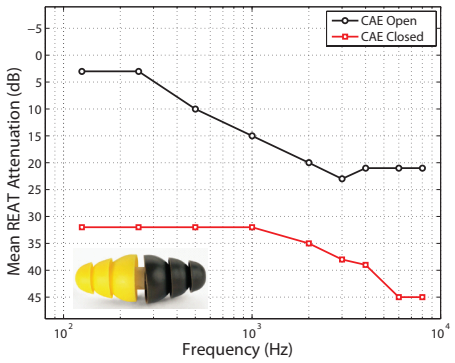


Figure 1: *Overall attenuation of Combat Arms Earplug in the Open and Closed configurations, as measured by the Real Ear Attenuation at Threshold (REAT) method. Adapted from Berger and Hamery [3].*

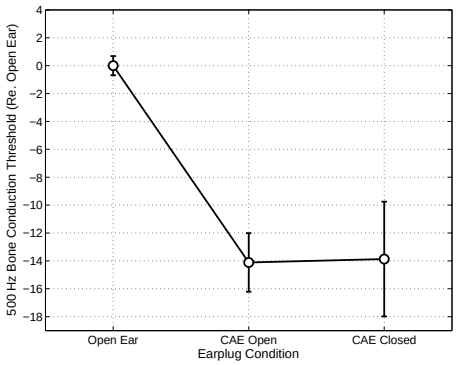


Figure 2: *Occlusion effect as measured by the difference in the 500 Hz bone conduction threshold with and without the earplug inserted in the ear. The results shown have been averaged across a total of ten subjects (seven males, three females). Error bars show the 95% confidence intervals in each condition.*

(e.g., gunfire) while allowing them to hear low-level sounds in the surrounding environment. This is accomplished by penetrating the earplug with a specially designed venturi vent that creates turbulence to block the sound transmission path when it is exposed to an impulse noise with a peak pressure in excess of 110 dB SPL. However, the CAE provides no protection from continuous noise, even at high levels. Thus, the plug is packaged in a unique dual-plug design that has a solid green flanged plug on one end and a yellow flanged plug with the venturi vent on the other end. The solid green side is inserted into the ear in situations where the listener is being exposed to high-level continuous noise (CAE Closed) and the yellow vented side is inserted into the ear for protection from impulse noise (CAE Open).

Figure 1 shows the attenuation characteristics of the CAE earplug, as measured in the Real Ear Attenuation at Threshold Method, in which attenuation is determined by the difference in the absolute sound detection threshold with and without the hearing protection device for narrow-band noises presented in a quiet room with a diffuse sound field [3]. As would be expected, the CAE produces much less attenuation in the Open configuration than in the Closed configuration, especially at low frequencies.

Prior to conducting the speech production experiment, a preliminary experiment assessed the occlusion effect generated by the earplugs based on the detection of a 500 Hz bone-conducted sound, with and without the earplug. The experiment was performed in a quiet, sound treated audiometric sound booth, with a 50 dB HL noise masker presented through an insert headphone in the non-test ear and a bone conduction transducer (Radio-Ear B71) placed over the temporal bone in the same ear. The opposite (test) ear either remained open or was fit with one of the two CAEs. In each trial, the listener used a slider attached to a MIDI soundcard (RME Hammerfall) to continuously adjust the level of the 500 Hz tone between two levels, one just above and one just below the threshold of detection. This alternation was repeated until two consecutive above-threshold and two consecutive below-threshold measurements were within 2 dB of one another. Once this was achieved, a lighted button above the slider was illuminated to allow the listener either to continue adjusting the level of the tone or to press a button to indicate that they were satisfied with the threshold measurement. The threshold was estimated from the mean of all four endpoints. Each of nine listeners provided a total of

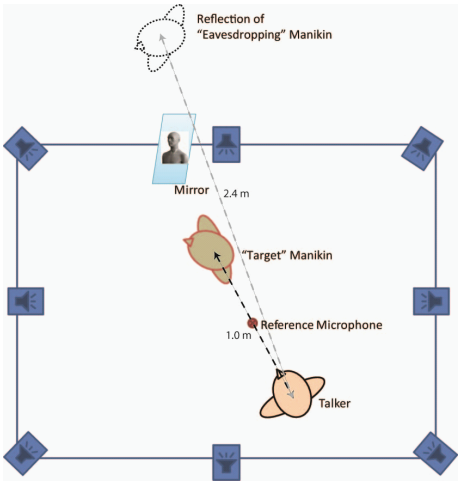


Figure 3: *Schematic diagram of room setup used to obtain the voice recordings.*

three estimates for each of the three earplug conditions in each ear.

The results of this preliminary experiment are shown in Figure 2. Both CAE earplugs produced a substantial occlusion effect that resulted in a reduction of approximately 14 dB in the 500 Hz bone conduction threshold relative to the open ear condition. Curiously, there was almost no difference in the 500 Hz detection thresholds between the two conditions, despite the substantially larger amount of attenuation in the CAE Closed condition at that frequency (32 dB vs 10 dB). Thus, it appears that the CAE Open earplug produces the same amount of occlusion but substantially less attenuation than the CAE Closed condition.

3. Speech Production Experiment

Once the occlusion effect for each earplug condition was determined, an experiment was conducted to determine the impact that the CAE earplugs had on speech production in noise. Talkers in a noise-filled sound booth were asked to privately communicate short verbal messages to a nearby acoustic manikin. The configuration of the testing room is shown in Figure 3. The

room itself was a standard sound booth, 2.75 by 2.6 m by 2.3 m, with a total of 16 loudspeakers (one centered on each wall and in each corner at ear level, and one in each upper and lower corner). For the purposes of this experiment, the talker (subject) was seated on a stool that was 1 m away from the location of a Knowles Acoustics Manikin for Acoustic Research (KEMAR) who was placed at the same height as the subject but was rotated away so the talker would be speaking at a point roughly 45° behind the manikin's right ear. This placement was used to replicate a situation where the talker had an urgent message that needed to be communicated verbally because the listener was facing in another direction and therefore could not be alerted to the message through hand gestures or other non-verbal means. In order to further constrain the talker's level of vocal effort, a mirror was placed behind the KEMAR in such a way that it reflected an image of a second manikin that appeared to be located 2.4 m away from the talker. This image was identified to the subjects as an eavesdropping listener, and they were instructed that their messages should be spoken loudly enough to be understood by the nearby target manikin, but not by the more distant eavesdropping manikin.

The experiment was divided into six 20-trial blocks, with each subject participating in two blocks with each type of hearing protection (open ear, CAE Open, or CAE Closed). At the start of each block, the subject was instructed to produce a 5-s quiet speech sample by reading a short passage about hurricanes. This speech sample was recorded both by the KEMAR in-ear microphones and by a reference microphone (Larson Davis 2527) located 0.5 m in front of the listener's mouth. Then, prior to the first trial of the experiment, the experimenter provided the subject with a number from 1 to 20 that corresponded to one of 20 quasi-militarily-relevant test phrases (e.g."Sgt. Kemar, we have a 10-91 in the Green Zone") printed on posters that were mounted on the wall of the sound booth behind the KEMAR manikin. Then the subject sat silently while a 4.5 s sample of noise was generated by all 16 loudspeakers in the room and recorded by the in-room microphones. This noise sample was a diffuse noise filtered to produce a pink noise spectrum at one of four different sound levels (50, 60, 70 or 80 dB SPL) at the location of the reference microphone. After the completion of this noise interval, there was a short pause, and then the exact same 4.5 s frozen noise sample was played through the loudspeakers in the room a second time. During this second interval, the subject read the designated target phase at a level that would be understandable by the nearby manikin but difficult to understand for the more distant eavesdropping manikin. This second speech-plus-noise sample was also recorded; the initial noise-only recordings were digitally subtracted from the speech-plus-noise recordings to recover a noise-free estimate of the speech signal at the locations of the three microphones in the room. Then the subjects were asked to give a verbal estimate of the amount of vocal effort used to produce the speech on a numerical scale ranging from 1 (not at all effortful) to 10 (extremely effortful).

The extracted speech waveforms were played and displayed to the experimenter, who used a mouse to manually extract the target phrase from the full 4.5 s recording. Recordings that were invalid (for example, because the subject made noise during the initial reference noise recording) were discarded, and randomly re-collected at some point before the end of the block. This procedure was repeated until a total of 5 speech samples were collected at each of the four noise levels.

Figure 4 displays examples of the magnitude spectra of the speech signals recorded during the experiment. The top row shows a case where the noise was set to 80 dBA, and the bottom row shows a case where the noise was set to 50 dBA. In both cases, the left panel displays the recording of the noise made in the first interval of each trial when the subject was sitting quietly, and the middle panel shows the recording in the second interval when the subject was asked to speak over the noise. The right panel replots these two curves, along with a third curve (solid black line) showing the difference between the two samples. This effectively represents a clean representation of the target speech signal. The gray line is the noise floor in each condition, calculated from the difference between two frozen noise samples when no speech signal was present in the second interval. Note that, in the 80 dB case, the technique is capable of recovering speech signals well below the level of the noise, and that, even at high frequencies (8 kHz), the speech recordings were well above the inherent noise floor of the waveform subtraction technique used to collect the speech samples.

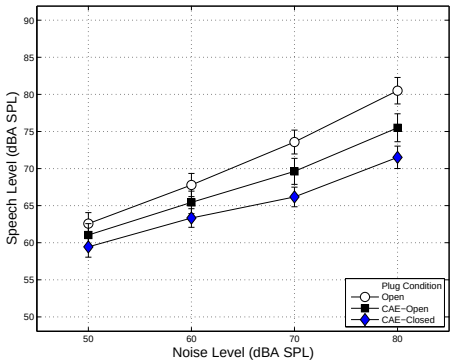


Figure 5: *Overall speech intensity as a function of room noise. The results have been averaged across 10 subjects with normal hearing (5 male, 5 female) ranging in age from 26 to 55 years (mean 36.8). The error bars show the 95% confidence intervals around each data point.*

4. Results

Figure 5 illustrates the relationship between the overall level of the speech samples produced by the subjects (as measured by the reference microphone 0.5 m away from the listener) and the overall level of the noise in the room. A few things are notable in this figure. First, it is apparent that, even in the open ear condition, the subjects generally did not increase their vocal intensity enough to maintain a constant signal-to-noise ratio in the high-noise conditions of the experiment. Second, it is apparent that both hearing protection conditions caused the listeners to reduce the intensity of their voices relative to the open-ear condition. In the 80 dB noise condition, subjects wearing the CAE Open earplugs spoke 5.8 dB less intensely than they did when they wore no hearing protection, and in the CAE Closed condition they spoke 10.1 dB less intensely than they did with no hearing protection. Notably, this 10 dB reduction in the CAE Closed is almost exactly the same as that reported by Tufts and Franks [2] for very similar flanged earplugs.

A more relevant estimate of the impact the hearing protectors had on speech communication in noise is shown in Figure 6, which shows the predicted speech intelligibility of the speech signals at the location of the manikin's better ear (the one oriented toward the talker) as a function of noise level and hearing protection condition. These estimates were obtained with the

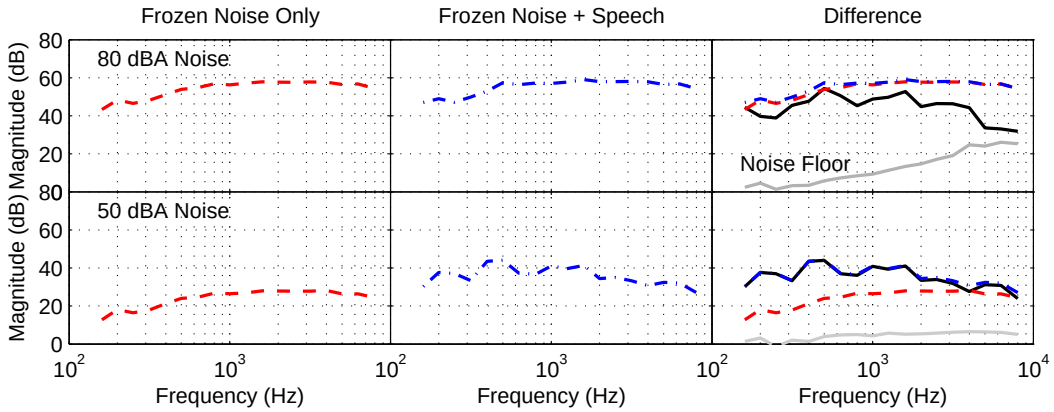


Figure 4: *Examples of magnitude spectra of speech and noise samples recorded during the experiment. The samples were collected using the pa_wavplayrecord function (written by Matt Frear and available at the MATLAB file exchange), which allows signals to be played and recorded synchronously through an ASIO soundcard with very little variation in time alignment across measurements. The gray lines in the right panel show the noise floor of the recording technique, which was estimated by subtracting two consecutively recorded samples of the noise waveform when the talker did not speak during either measurement interval.*

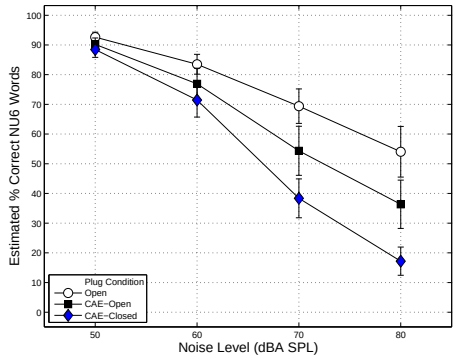


Figure 6: *Estimated speech intelligibility for NU6 words at the KEMAR ear closer to the talker as a function of noise level and hearing protector type.*

noise and speech measurements made from the in-ear microphone to calculate the Articulation Index (AI) in each condition [4], and then convert the AI score into an estimated percent intelligibility for NU6 words. In this figure, we can see that the subjects clearly did not increase their vocal intensity enough to maintain a high level of intelligibility when the noise level was above 60 dB, even in the open ear condition. The use of hearing protection in these high noise conditions had a devastating impact on predicted intelligibility in the 80 dB noise environment, with predicted NU6 scores dropping from 55% correct to 38% correct with the CAE Open earplugs, and plummeting to 18% with the CAE closed earplugs.

5. Conclusions

In this study, we examined the impact that hearing protection devices have on the production of speech in noisy environments. This study contributes to extant literature on this topic by specifically examining the impact of attenuation and occlusion for passive non-linear earplugs such as the CAE, which are often described during training as providing protection from high levels of impulse noise while being "acoustically transparent" in other low-level noise environments. Although these non-linear earplugs do produce relatively low attenuation, especially at low frequencies, the results of this experiment demonstrate that they actually produce just as much occlusion as a normal hearing protection device. When individuals wear these devices in environments that are noisy but not necessarily hazardous (i.e. $<$ 85 dBA), it is highly probable that these earplugs will bias them to speak more quietly than necessary to achieve effective communication. Therefore, training may be needed to make talkers aware of this potential issue and ensure that they do not incorrectly attribute this effect to some problem with the attenuation characteristics of the device. These results also indicate that the measurement technique described here, which uses waveform subtraction to recover the speech waveform produced by a talker in a noisy environment, may have many advantages over other methods of examining the effect of background noise on speech production. This technique could have many potential applications in future studies in this important area.

6. Acknowledgments

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Army, or Air Force, the Department of Defense, nor the U.S. Government

7. References

[1] F. Kuk, D. Keenan, and C.-C. Lau, "Vent configurations on subjective and objective occlusion effect," *Journal of the American Academy of Audiology*, vol. 16, no. 9, pp. 747–762, 2005.

[2] J. B. Tufts and T. Frank, "Speech production in noise with and without hearing protection," *The Journal of the Acoustical Society of America*, vol. 114, no. 2, pp. 1069–1080, 2003.

[3] P. Berger, E.H. Hamery, "Empirical evaluation using impulse noise of the level-dependency of various passive earplug designs," *Proceedings of Acoustics'08, Paris, France, June 29-July 4, 2008*, pp. 3719–3724, 2008.

[4] K. Kryter, "Methods for calculation and use of the articulation index," *Journal of the Acoustical Society of America*, vol. 34, pp. 1689–1697, 1962.

Compensation for Manipulated Auditory Feedback in Children with Specific Language Impairment

Michaela Hamel[1,3], Lisa Archibald[2], David Purcell[2,3]

[1]Graduate Program in Neuroscience, Western University, London, ON, Canada
[2]School of Communication Sciences & Disorders, Western University, London, ON, Canada
[3]National Centre for Audiology, Faculty of Health Sciences, Western University, London, ON, Canada
eholmes6@uwo.ca, larchiba@uwo.ca, purcelld@nca.uwo.ca

Abstract

Using a real-time vowel formant filtering system, F1 of children's productions of the word "head" /hɛd/ were shifted in real time and played to participants over headphones. This manipulated auditory feedback was shifted +340 Hz (/ɛ/ → /æ/) in one condition and -230 Hz (/ɛ/ → /ɪ/) in a second condition. Total compensation for this manipulated auditory feedback was monitored. On average, compensation occurred to oppose the manipulation. Children with Specific Language Impairment compensated significantly more than typically developing children.
Index Terms: auditory feedback, shifted formants, Specific Language Impairment.

1. Introduction

Auditory feedback is a mechanism by which an individual controls the characteristics of their voice, such as intensity, frequency, or speed. This mechanism is complex and can induce changes in shape or movement of speech-motor articulators that result in changes in speech characteristics. This process is a subconscious effort to match actual vocal outcomes with predicted vocal outcomes.

Auditory feedback has been described as an auditory analogue of reaching and grasping tasks. In tasks where individuals reached for objects while wearing vision-altering equipment, Gonzalez-Alvarez et al. observed that accuracy decreased when compared to baseline reaching accuracy without vision alteration [1]. Ma-Wyatt and McKee determined that when an object was moved during the reach, participants tended to reach for a location in between the original and final locations of the object [2]. Similarly, previous studies examining auditory feedback have found partial compensation for manipulated auditory feedback [3, 4].

The use of auditory feedback matures and changes over the human lifespan. In general, children older than 4 years, as well as adolescents and adults, respond to manipulated auditory feedback similarly, whereas on average children under the age of 4 years exhibit a negligible response [5]. Children with absent or decreased auditory feedback, such as those with prelingual deafness, have exhibited delayed grammar acquisition [6]. Individuals tend to compensate such that they oppose manipulations introduced into their auditory feedback. Individual variability exists in compensation for manipulated auditory feedback, where some individuals compensate completely and others not at all. On average, adults tend to compensate by less than 30% of the total manipulation introduced into their auditory feedback for the vowel /ɛ/.

Specific language impairment (SLI) refers to an unexpected delay in the development of oral language. In SLI, below average scores on language tests occur in spite of normal pure-tone thresholds and average scores on intelligence tests. Approximately 6-10% of the population is affected by this common impairment, and it is three times more likely to occur in males than in females [7, 8, 9].

Auditory feedback is thought to be important in the vocal refinement required for learning speech. SLI has been found to be related to decreased auditory processing abilities in highly specific tasks that rely on auditory pathways [10, 11, 12]. To investigate the relationship between the auditory feedback pathway and SLI, auditory feedback of the vowel /ɛ/ was manipulated and the resulting compensation for children with typical development (TD) or SLI was compared. It was hypothesized that children with SLI would respond differently to formant shifted auditory feedback than their TD peers.

2. Methods

Adult auditory feedback has been manipulated ±200 Hz in previous studies [3, 4]. To determine an appropriate manipulation for the London, Ontario child population, the words /hɛd/, /hæd/ and /hɪd/ were recorded six times each from 21 TD children aged 6-11 years in the London, Ontario school districts. These vowels were chosen since the vowel /ɛ/ was to be manipulated during the study. /ɛ/ has limited somatosensory feedback compared to point vowels and adults respond robustly to manipulation of its auditory feedback. Vowels /æ/ and /ɪ/ are in close proximity to /ɛ/ in formant space. The utterances containing /ɛ/, /æ/ and /ɪ/ were segmented to obtain the vowels. Formant frequencies were determined for each vowel and the average first (F1) and second (F2) formants were examined. The mean F1 distance /ɛ/ → /æ/ was +340 Hz and the distance /ɛ/ → /ɪ/ was -230 Hz. These manipulations

were used for the positive and negative shift conditions in this study.

Participants completed a perceptual frequency discrimination task that determined the smallest change in F1 that they could detect. Three words were presented over headphones with an accompanying animation on a computer monitor. The middle word was used as a reference point and always presented as /hɛd/. Participants selected which of two options sounded the most similar to the reference point. Initially, the F1 frequency difference was such that the two choices were /hɛd/ and /hæd/, making it simple for children to match the /hɛd/ option with the /hɛd/ reference point. Following this, the task used an adaptive two-alternative forced choice format in which the difficulty increased as participants made correct choices. This determined the minimum F1 difference participants needed to differentiate two /ɛ/-like vowels. This was completed to determine whether the minimum F1 difference participants needed to differentiate two /ɛ/-like vowels perceptually differed between the TD and SLI groups in this study.

In the altered auditory feedback task a real-time formant filtering method was used to alter F1 that children heard over headphones in two separate conditions. In the first condition, F1 was shifted +340 Hz from baseline (/ɛ/ → /æ/). After a short break during which the children were engaged in conversation, the second condition involving shifting F1 -230 Hz from baseline (/ɛ/ → /ɪ/) was completed. In both conditions F1 was shifted gradually by 10 Hz per production.

Participants (see Table 1) included 10 children (6-11 years old) with SLI and 10 TD children matched for gender, age, linguistic variables (Ontario English as the first and only language spoken at home), socioeconomic status, and nonverbal intelligence. All participants completed standardized tests of language (Composite Language Score of the CELF-IV [13]) and nonverbal intelligence (Performance IQ of WASI [14]) at each of two testing periods occurring one year apart as part of a larger study. Children with SLI scored more than one standard deviation (SD) below the mean (<85) and those with TD scored above this (>85) on the CLS at the final testing. Every effort was made to select children who showed a stable profile across the two testing periods, with resulting CLS discrepancy being less than 6 on average (SLI: $\mu_{(discrep)}$=7.4, SD=4.9; TD: $\mu_{(discrep)}$=3.8, SD=4.1; Both Groups: $\mu_{(discrep)}$=5.6, SD=4.8). Children whose behaviour was not conducive to completing the study tasks were not included in the reported sample (1 TD child).

A pure-tone audiometric hearing assessment was performed for both ears at octave frequencies between 250 and 4000 Hz. All participants had hearing thresholds 25 dB HL or better for all frequencies in both ears.

Participants wore a Shure WH20 headset microphone and Sennheiser HD 265 headphones, with the formant shift being introduced in real-time by National Instruments real-time hardware and custom software [3]. Participants sat in a sound-attenuated booth and were prompted via computer display. Six tokens were recorded for each of the vowels /ɛ/, /æ/ and /ɪ/. The most stable model order for an iterative linear predictive coding (LPC) formant tracker was determined for /ɛ/.

There were five phases in each condition. The first phase (15 trials) was used to allow participants time to become comfortable using the headphones and was not used in the analysis. In the second phase a baseline for the vowel /ɛ/ was established (20 trials). In the third phase, the "ramp", participant's auditory feedback was shifted 10 Hz per production, for a maximum manipulation of +340 Hz (34 trials) and -230 Hz (23 trials) from baseline in the positive and negative conditions respectively. In the fourth phase, the "hold", this maximum shift was maintained (20 trials). In the fifth phase, manipulation was removed and participants heard their unaltered voice (40 trials).

In offline analysis, the vowel portion of each production was segmented from neighbouring consonants using a semi-automated procedure and re-checked individually. A single value of F1, F2, and F3 was determined for each utterance by averaging formant estimates from the middle 80% of the vowel. F1 values were averaged across individuals for each trial in each group (SLI and TD). These average trials were then normalized by subtracting the average baseline for each group. Compensation was the change from baseline and was given by the normalized F1 in each trial. The Sign test was used to determine differences between the two groups' compensation for the hold trials where the maximum shift was employed. The Sign test is a non-parametric test with few assumptions that can be applied to small samples. Percent compensation was calculated by dividing the mean compensation during the hold phase by the shift size.

3. Results

Figure 1 displays the results of the F1 perceptual frequency discrimination task. Figures 2 and 3 display the average normalized F1 produced by the SLI and TD groups in the +340 Hz (positive manipulation) and -230 Hz (negative manipulation) conditions of the altered auditory feedback task, respectively. The average baseline and hold phases, as well as average percent compensation for the manipulations, are presented in Table 2. Table 3 displays individual average percent compensation for the positive shift (+340 Hz) condition. Both groups displayed similar distributions of percent compensation for the manipulation, with the SLI group displaying a higher percent compensation overall.

The ability of TD children and children with SLI to discriminate F1 of two /ɛ/-like vowels perceptually did not differ significantly. Compensation for the altered feedback manipulation was significantly greater for the SLI group than for the TD group for the positive +340 Hz condition ($p<0.005$). The difference between the two groups in the negative -230 Hz condition did not reach significance.

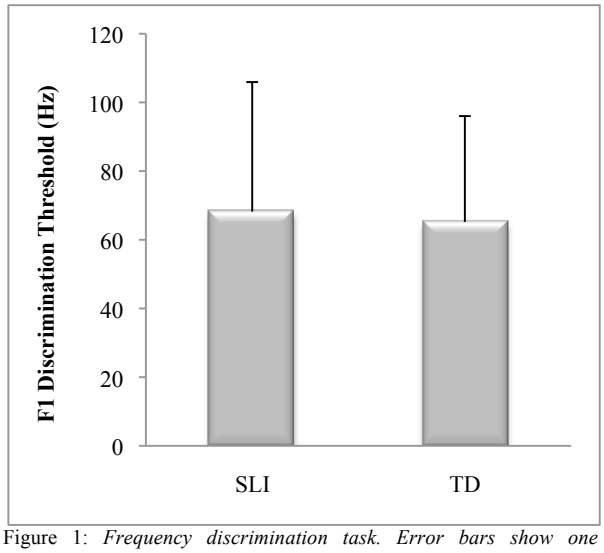


Figure 1: *Frequency discrimination task. Error bars show one standard deviation.*

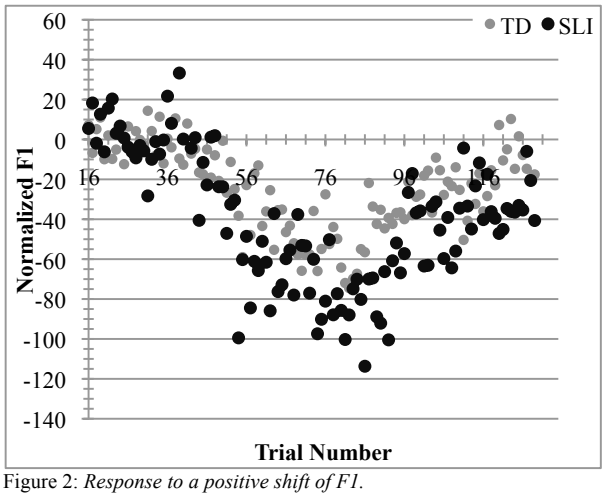


Figure 2: *Response to a positive shift of F1.*

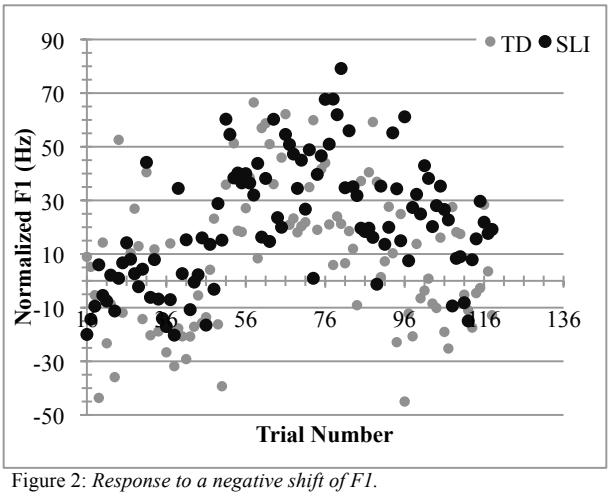


Figure 2: *Response to a negative shift of F1.*

Table 1: *Descriptive statistics for the SLI and TD groups. A statistically significant difference between SLI and TD group averages is indicated by ** (p<0.005).*

	Boys: Girls	Mean Age (Yrs)	S.D. Age (Yrs)	Mean CLS	S.D. CLS	Mean PIQ	S.D. PIQ
SLI	7:3	9.95	1.44	72**	7	93	12
TD	7:3	9.42	1.81	106**	14	96	5

Table 2: *Average F1 results for SLI and TD groups. A statistically significant difference in group average normalized F1 values in the hold phase between the SLI and TD groups is indicated by * (p<0.05).*

Group	Stimulus Manip. (Hz)	% Comp.	Baseline (Hz)	Hold (Hz)	Change (Hz)
SLI	+340	23	781	691	-89*
TD	+340	16	783	734	-48*
SLI	-230	12	780	807	+27
TD	-230	15	782	818	+35

Table 3: *Individual percent compensation for formant shifted auditory feedback in the TD and SLI groups.*

SLI Group	% Compensation	TD Group	% Compensation
Jafar	40	Alice	29
Lucky	35	Minnie	26
Eeyore	33	Aladdin	18
Rapunzel	31	Basil	18
Chip	31	Flower	16
Toby	26	Wilbur	15
Kanga	20	Hercules	14
Goofy	17	Merlin	12
Robin	3	Percy	5
Nana	-4	Stitch	4
Group Avg.	23	Group Avg.	16

4. Discussion

The performance of groups of TD children and those with SLI was compared on perceptual frequency discrimination and altered auditory feedback tasks. No difference between groups was found in perceptual frequency discrimination. In the altered auditory feedback task, baseline was similar for the vowel /ɛ/ across both groups and both the positive and negative conditions. Percent compensation for the two groups differed significantly in the positive condition where the SLI group exhibited more compensation than the TD group. There may be several reasons for this difference. Potentially, children with SLI may be more sensitive to manipulations in the auditory environment than typically developing children for certain stimulus conditions. This may mean that children with SLI are making different use of auditory feedback than typically

developing children. Other studies examining the auditory processing component of SLI using decision-making and forced-choice tests have found such differences, where children with SLI resolve frequency or respond dissimilarly to temporal information differently than TD children [14, 15]. Possibly, children with SLI may use the feedback they receive from auditory sources, including their own voices, in an atypical manner, perhaps being more sensitive to, or distracted by, the auditory environment than their TD peers. How this relates to their decreased language learning abilities is not yet understood.

The underlying reason why compensation was not significantly different in the negative shift condition is not known. These shift sizes were selected based upon several productions of /hɛd/, /hæd/ and /hɪd/ by a random selection of TD children from the London, Ontario school districts in order to ensure an appropriate shift for the London, Ontario child population. Analysis of the vowel spaces of the children in the present study indicates that this shift was more than sufficient to cross the average vowel boundaries of the children in the study. The vowel spaces of the SLI and TD groups were comparable. This seems to indicate that these were appropriate shift sizes to observe compensation, on average, in participants. However, the shift of -230 Hz may have been too small to sufficiently tease apart the two groups. Perhaps having a shift of -340 Hz, which would have matched the positive shift in magnitude, would have been sufficiently large to observe the difference between the SLI and TD subjects.

This study follows the previously noted result that children older than 4 years compensate on average for auditory feedback by opposing the manipulation [5]. No substantial changes were observed in F2 or F3 in response to the manipulation, as has been observed previously [16]. Further research is required to determine whether a negative shift of greater magnitude will better illustrate the difference in compensation between the two groups, as in the positive shift condition. The present study shows that children with SLI compensate more for manipulated auditory feedback than do their typically developing peers for certain stimulus conditions. Future research examining the nature of this compensation may reveal how the greater compensation observed for manipulated auditory feedback in these speech conditions is related to the development of language in children with SLI.

5. Acknowledgements

This research was funded by the Ontario Early Researcher Award and the Canadian Natural Sciences and Engineering Research Council (NSERC). The authors would like to thank Takashi Mitsuya, Dr. Ewen MacDonald, Dr. Ruth Martin, Dr. Marc Joanisse, Dr. Susanne Schmid, Dr. Janis Oram-Cardy and Allison Partridge for their help and support.

6. References

[1] Gonzalez-Alvarez, C., Subramanian, A. and Pardhan, S., “Reaching and grasping with restricted peripheral vision”, Opthamalic & Physiological Optics, 27(3):265-274, 2007.

[2] Ma-Wyatt, A. and McKee, S., “Visual information throughout a reach determines endpoint precision”, Experimental Brain Research, 179(1):55-64, 2006.

[3] Purcell, D. and Munhall, K., “Compensation following real-time manipulation of formants in isolated vowels”, Journal of the Acoustical Society of America, 119(4):2288-2297, 2006.

[4] Purcell, D. and Munhall, K., “Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation”, Journal of the Acoustical Society of America, 120(2):966-977, 2006.

[5] MacDonald, E., Johnson, E., Forsythe, J., Plante, P. and Munhall, K., “Children’s development of self-regulation in speech production”, Current Biology, 22(2):113-117, 2011.

[6] Nikilopoulos, T., Dyar, D., Archbold, S., and O’Donoghue, G., “Development of spoken language grammar following cochlear implantation in prelingually deaf children”, Archives of Otolaryngology – Head & Neck Surgery, 130(5):629-633, 2004.

[7] Tomblin, J., Records, N., Buckswalter, P., Zhang, X., Smith, E. and O’Brien, M., “Prevalence of Specific Learning Impairment in kindergarten children”, Journal of Speech, Language and Hearing Research, 40:1245-1260, 1997.

[8] Bonneau, D., Verny, C., and Uze, J., “Genetics of specific language impairments”, Archives of Pediatrics, 11(10):1213-1216, 2004.

[9] Bishop, D., “Genetic and environmental risks for specific language impairment in children”, Philosophical Transactions of the Royal Society: Biological Sciences, 356(1407):369-380, 2001.

[10] Goffman, L., “Prosodic influences on speech production in children with Specific Language Impairment and Speech Deficits: Kinematic, acoustic and transcription evidence”, Journal of Speech, Language and Hearing Research, 42:1499-1517, 1999.

[11] McArthur, G. and Bishop, D., “Speech and non-speech processing in people with specific language impairment: A behavioural and electrophysiological study”, Brain and Language, 94:260-273, 2005.

[12] Semel, E., Wiig, E., Secord, W., “Clinical Evaluation of Language Fundamentals – 4”, San Antonio, TX: Psychological Corp/Harcourt, 2003.

[13] Wechsler, D., “Wechsler Adult Intelligence Scale – 3: Administration and scoring manual”, San Antonio, TX: Psychological Corp, 1997.

[14] Miller, C. and Wagstaff, D., “Behavioral profiles associated with auditory processing disorder and specific language impairment”, Journal of Communication Disorders, 44(6):745-763, 2011.

[15] Wright, B., Lombardino, L., King, W., Puranik, C., Leonard, C. and Merzenich, M., “Deficits in auditory temporal and spectral resolution in language-impaired children”, Nature, 387:176-178, 1997.

[16] MacDonald, E., Purcell, D., and Munhall, K., “Probing the independence of formant control using altered auditory feedback”, Journal of the Acoustical Society of America, 129(2):955-965, 2011.

CHARACTERIZING PHONETIC CONVERGENCE WITH SPEAKER RECOGNITION TECHNIQUES

Amélie Lelong and Gérard Bailly

GIPSA-lab, UMR 5216 CNRS/INPG/UJF/U. Stendhal
{amelie.lelong,gerard.bailly}@gipsa-lab.grenoble-inp.fr

Abstract

Speakers are known to accommodate to each other's behavior when interacting. Information circulates via multiple sensory-motor loops operating at various levels of the interaction and this closed-loop process induces modifications in all levels of representation from social and psychological evaluation to low-level gestural behaviors such as gaze, respiratory patterns, or speech. Various authors have proposed that these representations tend to converge or diverge according to cognitive demand. While quite plausible, claimed observations of such behavior in speech are extremely controversial. The effects are rather small, and are difficult to capture and characterize objectively. This paper focuses on the study of the convergence between phonetic representations – spectral realizations of speech sounds – using automatic classification techniques developed for speech and speaker recognition. Using data collected during a novel language game we term 'verbal dominoes', we show that scores are comparable between global techniques and a more fine-grained analysis focused on vocalic segments.

Index Terms: speaker recognition, phonetic convergence, speech adaptation

1. Introduction

Individuals accommodate their communication behavior [1] either by becoming increasing similarity with their interlocutors (i.e. convergence) or on the contrary by increasing their differences (i.e. divergence). Speech accommodation has been observed at several levels. Researchers have in fact conducted studies on convergence of phonetic dimensions such as pitch, speech rate, loudness or dispersions of vocalic targets. The supposed goals and benefits of this adaptation include: easing comprehension, facilitating the exchange of highly context-dependent messages, disclosing ability and willingness to perceive, understanding or accepting new information, and maintaining social glue. Zoltan-Ford [2] has also shown that users of dialog systems tend to converge lexically and syntactically to the spoken responses of the system. Moreover, Ward et al [3] demonstrated that adaptive systems mimicking this behavior facilitate learning. But the phenomenon depends on several factors and most objective studies show only limited convergence, if any.

This emerging field of research is nonetheless central for two projects: the study of adaptive behavior during unconstrained conversation, and the substitution of an artificial conversational agent for a live partner.

This paper addresses two main topics: (a) we document a new method of collecting phonetic material to study and isolate the impact of the various factors influencing adaptation; and (b) we evaluate automatic techniques for quantifying any extant degree of convergence.

2. Observing and characterizing phonetic convergence

2.1. Scenarios

Researchers have used a variety of paradigms to characterize adaptation at different levels.

Perturbation of auditory feedback: Evidence that speakers tend to compensate for perturbation of their auditory feedback (see [4] for f0) lead some researchers to infer an *internal* sensory-motor speech representation towards which speakers tend to return in response to *external* excitations (or in their absence).

Imitation: Repetition and shadowing paradigms demonstrate convergence effects on Voice Onset Time (VOT) [5], F0 distribution [6], and articulation [7]. Sato et al. [8] showed that unintentional and voluntary imitation during the production of vowels used almost the same cognitive resources and resulted in similar behavior.

Ambient production: Delvaux and Soquet [9] tested the influence of ambient speech on pronunciation of certain keywords during a description task. They show small but significant effects on the spectra of target sounds when uttered in alternation with recordings of same vs. different Belgian dialects of French.

Interaction: Finally, researchers have studied phonetic convergence during unconstrained interaction. Pardo [10] examined the pronunciation of target words exchanged during a map task between pairs of same-sex talkers. Her perceptual experiments show that interactive speech decreases inter-subject distances. Aubanel and Nguyen [11] tested a new method of collecting dense interactive corpora with uncommon proper nouns, and they found a number of significant signatures of dialectal and phonetic convergence.

2.2. Computing degree of convergence

Quantification of convergence requires a baseline for comparison, so the default phonetic characteristics of speech segments (words, syllables, allophones) that will be analyzed during interaction are thus often collected through reading [10-11] or playing games alone [9] in a so-called pre-test session. Phonetic characteristics of the two speaker's productions before interaction are then compared to those of speech segments uttered during the interaction or after (post-test).

To characterize phonetic convergence, most authors use spectral cues (formants, Mel Frequency Cepstral Coefficients ie. MFCC) in the central part of particular segments of target words (mostly vowels & fricatives). The calculation of RMS distances between speaker-specific allophones are sometimes preceded by linear discriminant analysis [11].

To our knowledge, no results have been published addressing more global acoustic characterizations.

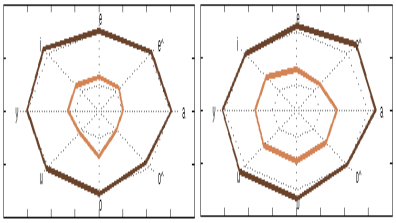


Figure 1. *Convergence rates C_{LDA} for the 8 vowels exchanged by two dyads. Circles with dotted lines (radius 1 and 4) feature default vocalic representations of the two speakers. Left: no convergence was found; Right: convergence of one speaker (C_{LDA}=0.38) towards the other, the orange line is getting closer to the outer circle representing the default vocalic representation of the partner.*

3. Speech recognition

The method for calculating convergence rates used by Delvaux & Soquet [9] and numerous researchers consists in computing an average distance between vocalic spaces using linear discriminant analysis (LDA). MFCC are first extracted for each vocalic target. Discriminant spaces are then computed for the central frames of each target sound for each pair of speakers. These frames are finally projected on the first discriminant axis separating speaker-specific spaces for the pronounced vowel and convergence rates for each target sound are calculated by normalizing the distance between speakers during interaction by the distance between vowels uttered during the pre-test. The convergence rate - noted as C_{LDA} - is then taken as the mean of these sound-specific rates (see Figure 1).

This method requires prior segmentation, labeling and clustering of specific target sounds (here vowels), the pronunciation of which speakers are supposed to mutually accommodate.

4. Speaker recognition

This paper compares the previous approach with a speaker recognition technique that compares the more global shape of the acoustic spaces. The experiments were performed using the MISTRAL platform [12]. We choose to perform speaker recognition by the Gaussian mixtures models (GMM), one of the most popular techniques for text-independent speaker recognition [13]. The speaker decision task mainly consists in a basic statistical test between two hypotheses:

- H_S: the speech characteristics *y* has been produced by the hypothesized speaker S
- $H_{\neg S}$: *y* is not from the hypothesized speaker S (often called the model of the "world")

The decision uses a likelihood ratio (LR_S) test given by:

$$LR_S(Y) = \prod_{y \in Y} \frac{p(y \mid H_S)}{p(y \mid H_{\neg S})} < \theta \qquad (1)$$

where $p(y|H)$ is the probability density function for the hypothesis H evaluated for the speech segment y and θ is the decision threshold for accepting or rejecting H_S.

With MISTRAL, the log likelihood ratio (*LLR*) is computed over a test set of frames *Y*. Two GMM respectively describe $p(y|H_S)$ and $p(y|H_{\neg S})$ with the following law:

$$p(y \mid H) = \sum_{i=1..M} w_i N(y \mid \mu_i, \Sigma_i) \qquad (2)$$

where wi, μi and Σi are the weights, mean vectors and covariance matrix of the M components of the mixture.

In our case, H_S and $H_{\neg S}$ are the models of the two speakers of the dyad: the "world" $\neg S$ corresponds to the interlocutor's model. We then note:

$$LLR_{l1l2}(Y) = \sum_{y \in Y} \log\left(\frac{p(y \mid H_{l1})}{p(y \mid H_{l2})}\right) \qquad (3)$$

GMMs here have M=64 components and the *y* components are MFCC coefficients estimated every 10ms.

These GMMs are trained in order to maximize $LLR_{s1s2}(P_{s1})+LLR_{s2s1}(P_{s2})$[(1)] over the set of frames P_{s1} and P_{s2} uttered respectively by speakers *s1* and *s2* during the pretest. This sum corresponds to the global distance between acoustic spaces of the two speakers.

Initialized using vector quantization, GMM parameters are refined by the iterative Expectation-Maximization (EM) algorithm in order to increase the likelihood of the estimated model for the observed feature vectors. Five to ten iterations are sufficient to get a correct estimation of each speaker's model.

The convergence rate of *s1* "towards" *s2*, noted $C_{LLR}(s1 \rightarrow s2)$ is then taken as the relative quotient between the difference of a speaker's LLR (here *s1*) calculated with his own model on frames P_{s1} and during interaction (I_{s1s2}) and the difference of LLR calculated with the two interlocutor's model on the pre-test (P_{s1}).

$$C_{LLR}(s1, s2) = \frac{LLR_{s1s2}(P_{s1}) - LLR_{s1s2}(I_{s1s2})}{LLR_{s1s2}(P_{s1}) - LLR_{s1s2}(P_{s2})} \qquad (4)$$

where I_{s1s2} is the set of frames uttered by speaker *s1* when interacting with speaker *s2*. So, if we don't have any convergence, $I_{s1s2}=P_{s1}$ and $C_{LLR}(s1 \rightarrow s2) = 0$.

5. Data

5.1. Speech dominos

A novel technique called "Speech Dominoes" [14] was used to collect rich phonetic data on interactive speech. The rule of the game is quite simple: speakers had to choose between several alternatives the word that begins with the same syllable as the final one of the word previously uttered by the interlocutor (see Figure 2). The experiment was divided into two phases. Intrinsic references were gathered for each speaker during a *pre-test* session, where the speaker reads aloud a list of 350 words before any interaction with others. The pre-test words are those used during the dominoes' game. During the game, each interlocutor pronounces respectively half of the *pre-test*

(1) Since $LLR_{s1s2}(P_{s1}) = - LLR_{s2s1}(P_{s1})$, $LLR_{s1s2}(P_{s1}) + LLR_{s2s1}(P_{s2}) = LLR_{s1s2}(P_{s1}) - LLR_{s1s2}(P_{s2})$

words, i.e. about 175 words. Figure 2 represents the first speech dominoes used in the interactive scenario. Interlocutors have to choose and utter alternatively the rhyming words. At each turn, speaker has to wait for his interlocutor to utter the correct word in order to choose what to pronounce next since the alternatives given to him are equally probable (e.g. both words [tɔrdy] and [tɔrʃi] exist in French and have roughly the same lexical frequency). Our reference subject first pronounces [rotɔr] to begin the game, then our tested subject will have to choose between [tɔrdy] and [bɛrly] the one that begins with [tɔr]: he will thus utter [tɔrdy] and so on.

We chain here simple dissyllabic words in order to limit the cognitive load and ease the running of successive sessions. As our first analyses are focused on vowels [15], we select bi-syllabic words chosen to collect equal numbers of allophonic variations (about 20 tokens per speaker) of the eight peripheral oral French vowels: [a], [ɛ], [e], [i], [y], [u], [o], [ɔ].

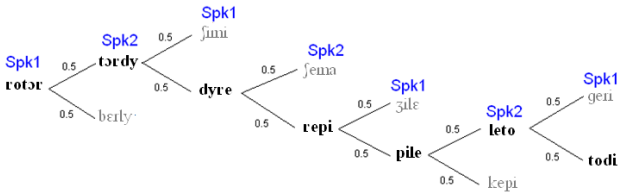


Figure 2. *First speech dominoes used in the interactive scenario. Correct rhymes in each pair are enlightened in bold.*

Table 1. *Convergence rates computed for each member of our 27 pairs by LDA (first column and second column of interact and pretest corresponds respectively to the mean convergence rate and its standard deviation) and LLR. Significant data (p<0.1 for LDA distributions for Interact vs. Pretest) are in bold.*

Game	Dyad	Initiator							Respondent						
		Sex	LDA				p	LLR	Sex	LDA				p	LLR
			Interact		Pretest					Interact		Pretest			
Session 1: 186 dominos - strangers	1	M	.03	.10	.03	.01	.89	.11	F	.03	.09	.04	.02	.79	.18
	2	M	.01	.05	.03	.02	.20	.14	F	.04	.10	.04	.02	.90	.01
	3	M	-.01	.09	.04	.02	.19	.20	F	.11	.14	.04	.02	.16	.09
	4	M	.04	.13	.09	.06	.32	.05	**M**	**.13**	**.08**	**.07**	**.03**	**.07**	**.08**
	5	M	.07	.14	.08	.05	.89	.25	**M**	**.28**	**.20**	**.07**	**.06**	**.01**	**-.05**
	6	M	.06	.19	.06	.03	.95	.16	**M**	**.31**	**.17**	**.05**	**.03**	**.00**	**.15**
	7	F	.01	.13	.09	.05	.14	.07	F	.15	.15	.08	.05	.21	.08
	8	F	.10	.19	.09	.06	.84	.05	M	.08	.11	.07	.06	.87	.07
	9	**F**	**.41**	**.38**	**.11**	**.06**	**.04**	**.30**	F	.18	.29	.08	.04	.33	.15
	10	M	.17	.24	.09	.08	.37	.19	M	.15	.11	.09	.09	.23	.06
	11	M	.08	.19	.07	.01	.78	.21	M	.04	.14	.07	.03	.46	.10
	12	M	.08	.09	.04	.02	.23	.09	**F**	**-.04**	**.07**	**.03**	**.02**	**.02**	**.07**
Session 2: 350 dominos - friends	13	**M**	**.41**	**.31**	**.07**	**.03**	**.01**	**.21**	M	.14	.12	.07	.02	.16	.18
	14	M	.15	.19	.05	.03	.16	.25	M	.13	.16	.04	.03	.16	.07
	15	**M**	**.40**	**.29**	**.07**	**.06**	**.01**	**.44**	M	.03	.22	.07	.05	.60	.16
	16	F	.00	.11	.03	.02	.49	.13	M	-.03	.12	.02	.01	.22	.11
	17	F	.00	.09	.03	.02	.50	.12	M	.04	.07	.03	.02	.46	.07
	18	F	.06	.12	.02	.01	.32	.13	**M**	**.10**	**.10**	**.02**	**.01**	**.04**	**.16**
	19	F	.14	.24	.06	.04	.38	.38	F	.13	.31	.08	.06	.66	.11
	20	**F**	**.39**	**.24**	**.06**	**.03**	**.00**	**.46**	F	.00	.14	.07	.02	.16	.08
	21	F	.16	.31	.04	.02	.30	.23	**F**	**.22**	**.25**	**.05**	**.03**	**.07**	**.13**
	22	F	.07	.33	.10	.07	.80	-.10	**F**	**.28**	**.20**	**.09**	**.06**	**.02**	**.16**
	23	**F**	**.15**	**.15**	**.06**	**.02**	**.09**	**.20**	**F**	**.18**	**.11**	**.06**	**.02**	**.01**	**.18**
	24	F	.22	.43	.08	.04	.36	.23	F	.34	.54	.09	.05	.21	.51
	25	F	.12	.16	.05	.03	.26	.28	**F**	**.15**	**.13**	**.06**	**.04**	**.08**	**.29**
	26	**F**	**.12**	**.14**	**.01**	**.01**	**.05**	**.39**	M	-.03	.11	.01	.01	.32	-.03
	27	**F**	**.34**	**.24**	**.06**	**.02**	**.01**	**.48**	F	.22	.35	.06	.02	.23	.27

5.2. Speakers' models, reference and test data

Only half of the pre-test data are used to train the speakers' models. The other half is used as reference data. We used a simple cross validation procedure: the convergence rates are the mean values of relative distances between reference and test data over 10 random partitions between training and reference data.

In a first series of experiments with 186 dominos [15], we noted that phonetic convergence was higher for dyads who already knew each other and particularly for women [10], as shown in the first 12 pairs in Table 1 and Figure 3. During this condition, speakers were in two different rooms and communicated through microphones and headphones. This setup was easy to realize thanks to the MICAL platform of our laboratory (two rooms separated with a tinted mirror). In a second series involving good friends exchanging a larger number of dominoes (350), 3 male dyads, 4 mixed dyads and 8 female dyads have been tested. 5 men from 24 to 54 years old and 11 women from 18 to 26 years old participated. In this case, people were engaged in a real face-to-face interaction.

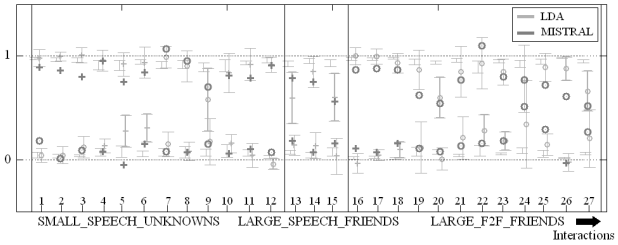


Figure 3. *Comparison of convergence rates for the two methods (1 is for the initiator and 0 for the respondent).The dark gray color represents the results obtained with MISTRAL. The results obtained with the different methods are quite similar. During interactions 1 to 15, people were communicated thanks to microphones and headphones while they were in a face-to-face condition for interactions 16 to 27.*

6. Results

Table 1 displays convergence rates C_{LDA} and C_{LLR} computed for all dyads. An ANOVA test was performed to test significant deviations between reference and test C_{LDA}. Distributions with significant convergence rates ($p<0.1$) are noted in bold.

Convergence is not systematic. Moreover, we can see that the phenomenon is amplified for same sex pairs (see pairs 4-6, 9, 13-15, 20-23 and 25-27) and particularly for women. This observation led us to select mostly women for the final interactions, and the results largely confirm this tendency.

We found a significant correlation of .64 ($p<0.05$) between these two coefficients for initiators and of .73 ($p<0.005$) for the respondents in the case of the large corpus (last 15 interactions in Table 1 and Figure 3). The correlations calculated on the first 12 interactions are lower. We do think that this is the consequence of insufficient training data provided by the 186 dominos.

6.1. Convergence and performance

We define turn-taking time (TTT) as the time delay between the onset of the last vowel of the domino pronounced by one speaker and the onset of the first vowel of the next domino uttered by his partner. Figure 4 shows the main impact of convergence on the evolution of TTT during the interaction: for moderate convergence rates, the degree of convergence of the initiator towards his partner correlates with increased TTT speed for the latter (r = -.77). We do not find this effect for the initiator's turns. This tends to confirm that the role and background of each participant has a strong impact on behavior and performance [16].

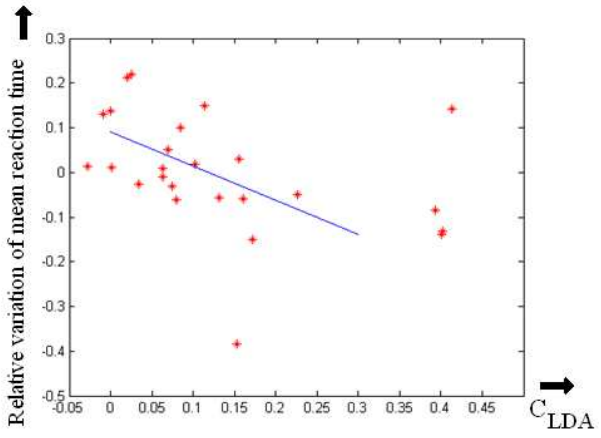


Figure 4. *Relative variation of mean turn-taking time of the respondent as a function of C_{LDA} of the initiator. Test subjects increase the rhythm of the interaction (decrease turn-taking time) in response of the phonetic convergence of their interlocutor.*

7. Conclusions

We have shown here that speaker recognition techniques provide a reliable estimate of the global degree of phonetic convergence without the need of phonetic segmentation or a procedure for part of speech pairing. For almost all pairs analyzed so far, few cases of divergence have been observed. On the contrary, large convergence rates have been found. Such impoverished phonetic contrasts between interacting speakers should be considered in automatic speaker tracking.

Our interaction paradigm offers other potential applications as well, regarding for instance the impact of word frequencies on the convergence [17] or rhythmical coupling across interlocutors. A perceptual validation of the large convergence effects found here is also called for.

This method will be used to characterize adaptation in less controlled conditions, to investigate the impact of conditions and linguistic content and study the dynamics of phonetic convergence. We plan to train statistical speech synthesis engines to implement the dynamics of the observed adaptation strategies. Such interlocutor-aware components are certainly crucial for creating social rapport between humans and virtual conversational agents [18].

8. Acknowledgements

This work was supported by ANR Amorces and by the Cluster RA ISLE. We thank Frederic Elisei, Sascha Fagel and Loïc Martin for their help.

9. References

[1] Giles, H., J. Coupland, and N. Coupland, *Contexts of Accommodation: Developments*. Applied Sociolinguistics. 1991, Cambridge: Cambridge University Press.

[2] Zoltan-Ford, E., *How to get people to say and type what computers can understand.* International Journal of Man-Machine Studies, 1991. **34**: p. 527-547.

[3] Ward, A. and D. Litman. *Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora*. in *SLaTE Workshop on Speech and Language Technology in Education*. 2007. Farmington, PA.

[4] Jones, J.A. and K.G. Munhall, *Perceptual calibration of F0 production: Evidence from feedback perturbation.* Journal of the Acoustical Society of America, 2000. **108**: p. 1246-1251.

[5] Fowler, C.A., et al., *Cross language phonetic influences on the speech of French-English bilinguals.* Journal of Phonetics, 2008. **36**(4): p. 649-663.

[6] Gregory, S.W., S. Webster, and G. Huang, *Voice pitch and amplitude convergence as a metric of quality in dyadic interviews.* Language and Communication, 1993. **13**: p. 195-217.

[7] Gentilucci, M. and P. Bernardis, *Imitation during phoneme production.* Neuropsychologia, 2007. **45**(3): p. 608-615.

[8] Sato, M., et al. *Converging to a common speech code: automatic imitative and perceptuo-motor recalibration processes in speech communication.* in *Second Neurobiology of Language Conference*. 2010. San Diego, USA.

[9] Delvaux, V. and A. Soquet, *The influence of ambient speech on adult speech productions through unintentional imitation.* Phonetica, 2007. **64**: p. 145-173.

[10] Pardo, J.S., *On phonetic convergence during conversational interaction.* Journal of the Acoustical Association of America, 2006. **119**(4): p. 2382–2393.

[11] Aubanel, V. and N. Nguyen, *Automatic recognition of regional phonological variation in conversational interaction.* Speech Communication, 2010. **52**: p. 577-586.

[12] Charton, E., et al. *Mistral: an open source biometric platform* in *25th Symposium on Applied Computing (SAC)*. 2010. Switzerland.

[13] Reynolds, D., *Speaker identification and verification using Gaussian mixture speaker models.* Speech Communication - special issue on Face-to-Face Communication, 1995. **17**(1): p. 91-108.

[14] Bailly, G. and A. Lelong. *Speech dominoes and phonetic convergence*. in *Interspeech*. 2010. Tokyo. p. 1153-1156.

[15] Lelong, A. and G. Bailly, *Study of the phenomenon of phonetic convergence thanks to speech dominoes* in *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issue*, A. Esposito, et al., Editors. 2011, Springer Verlag: Berlin. p. 280-293.

[16] Pardo, J.S., I. Cajori Jay, and R.M. Krauss, *Conversational role influences speech imitation.* Attention, Perception, & Psychophysics, 2010. **72**: p. 2254-2264.

[17] Goldinger, S.D., *Echoes of echoes? An episodic theory of lexical access.* Psychological Review, 1998. **105**: p. 251-279

[18] Gratch, J., et al. *Creating rapport with virtual agents*. in *Intelligent Virtual Agents (IVA)*. 2007. Paris, France. p. 125-138.

A Preliminary Study of Individual Responses to Real-Time Pitch and Formant Perturbations

Ewen N. MacDonald[1], *Kevin G. Munhall*[2]

[1]Centre for Applied Hearing Research, Technical University of Denmark, Denmark
[2]Department of Psychology, Queen's University, Canada

emcd@elektro.dtu.dk, kevin.munhall@queensu.ca

Abstract

Previous studies have demonstrated a wide range in individuals' compensations in response to real-time alterations of the auditory feedback of both pitch and formant frequencies. One potential source of this variability may be individual differences in the relative weighting of auditory and somatosensory feedback. The present study examined this variability by comparing individuals' compensations during two perturbation conditions: a pitch shift (+200 cents) and a formant shift (F1 +200 Hz, F2 -250 Hz). While no significant correlation was found between the two perturbation conditions, a modest correlation between compensations in pitch and formant frequency was observed within the pitch perturbation condition.

1. Introduction

When we talk, we monitor the sounds we produce to aid us in controlling speech production. Traditionally, this use of auditory feedback has been studied using perturbation techniques in which characteristics of the acoustic signal such as pitch [1] or formants [2] are altered in real-time. On average, talkers compensate by adjusting the acoustics of their voice in the direction opposite to that of the perturbation. However, the compensatory response has been found to vary significantly across individuals with some "following" rather than opposing the perturbation [3, 4]. For example, in a recent study that examined the compensation of 116 female talkers in response to the same formant shift (+200 Hz in F1, -250 Hz in F2), the average compensation (53 and 58 Hz in F1 and F2) was similar to the standard deviation of the compensations (44 and 69 Hz respectively) [5].

Somatosensory feedback also plays a role in speech-motor control [6]. Adapting speech to completely compensate for acoustic perturbations may result in somatosensory feedback that is incongruous. Evidence of increased compensation to pitch shifts when a local anesthetic was administered to the vocal folds supports this tradeoff between auditory and somatosensory feedback [7]. Thus, a possible explanation for the large variability in talkers' compensation may be individual differences in the relative weighting of auditory vs. somatosensory feedback in speech-motor control.

The approach taken in the present study was to investigate this relative weighting hypothesis by comparing the compensatory responses to pitch and formant shifts. A talker who relies mostly on auditory feedback should exhibit large compensations for both pitch and formant perturbations but a talker who relies mostly on somatosensory feedback should exhibit little compensation.

While pitch and formant perturbation experiments are similar conceptually, the experimental paradigms are often quite different. In a traditional pitch perturbation experiment, talkers are asked to produce sustained vowels, often with durations greater than 5 s. Over the course of a sustained vowel utterance, one or more short perturbations, with durations ranging from 100–500 ms, are randomly introduced. In contrast, in a typical formant perturbation experiment, talkers produce single-syllable utterances and when a perturbation is applied, it is applied over an entire utterance. In the present experiment, the paradigm typical of formant perturbation experiments was used for both the pitch and formant perturbation conditions. The main advantage in using this paradigm is that it allows us to examine if talkers exhibit pitch and formant compensation in the same set of utterances.

2. Method

2.1. Participants

The participants in this study were 22 undergraduate female students at Queen's University. All were native English speakers and reported no history of hearing or language disorders. All were found to have normal hearing thresholds between 500 and 4000 Hz (i.e., < 25 dB HL).

2.2. Equipment

The equipment used to conduct the real-time formant shifting was identical to that used by MacDonald et al. [8]. Testing was conducted in an Industrial Acoustics Co. (IAC) sound booth. Talkers spoke into a headset microphone (Shure WH20). Signal conditioning was performed using amplification (Tucker-Davis Technologies MA3 microphone amplifier), and low-pass filtering (cut-off frequency of 4500 Hz, Krohn-Hite 3384 filter).

For the condition where formants where shifted, the conditioned signal was digitized with a sampling rate of 10 kHz and filtered in real-time using custom software running on a National Instruments PXI-8106 controller. Formants were estimated every 900 µs using an iterative Burg algorithm with a model order that varied from 8 to 12 across individuals. IIR filter coefficients were computed based on these estimates such that a pair of spectral zeroes was placed at the location of the existing formant frequency and a pair of spectral poles was placed at the desired frequency of the new formant.

For the condition where the pitch was shifted, the conditioned signal was processed using an Eventide Harmonizer H3000 employing a proprietary algorithm.

The processed output was amplified and mixed with noise (Madsen Midimate 622 audiometer) and presented over headphones (Sennheiser HD 265) such that the speech and noise were presented at approximately 80 and 50 dBA, respectively.

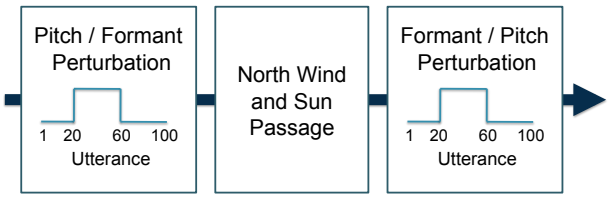


Figure 1: *Schematic of the procedure.*

2.3. Procedure

After collecting pure-tone hearing thresholds, talkers produced six utterances of seven English vowels in an /hVd/ context ("heed," "hid," "hayed," "head." "had," "hawed," and "who'd"). Talkers were instructed to say words that appeared on a computer monitor at a natural rate and speaking level. Each word prompt lasted 2.5 s and the inter-trial interval was approximately 1.5 s. These utterances were analyzed to select the best model order for each individual, using a heuristic based on minimum variance in formant frequency over a 25 ms segment midway through the vowel.

Each talker participated in a Pitch Perturbation condition and a Formant Perturbation condition. The order in which talkers completed the conditions was counterbalanced. Between the conditions, talkers read aloud "The North Wind and the Sun" passage [9]. An overall schematic of the experiment is illustrated in Figure 1.

In each of the perturbation conditions, talkers produced a total of 100 utterances of the word "head." For the first 20 utterances, the Baseline phase, talkers received normal auditory feedback (i.e., amplified and mixed with noise, but with no pitch or formant shift). For utterances 21–60, the Shift phase, talkers received altered auditory feedback. In the Pitch Perturbation condition, the auditory feedback was increased by 200 cents. In the Formant Perturbation condition, F1 was increased by 200 Hz and F2 was decreased by 250 Hz. For utterances 61–100, the Return phase, auditory feedback was returned to normal.

The procedure used for offline analysis was similar to that used by MacDonald et al. [8]. The boundaries of the vowel segment in each utterance were estimated using an automated process based on the harmonicity of the power spectrum. These boundaries were then inspected by hand and corrected, if required.

For each vowel segment, F0 estimates were calculated using Praat software (www.praat.org). A single "steady-state" value was calculated from the median of the estimates from 40% to 80% of the way through the vowel.

The first three formant frequencies were estimated offline from the first 25 ms of a vowel segment with the same algorithm used in the online shifting. The formants were estimated again after shifting the window 1 ms and repeated until the end of the vowel segment was reached. For each vowel segment, a single steady-state value for each formant was calculated by averaging the estimates for that formant from 40% to 80% of the way through the vowel. While using the best model order reduced gross errors in formant tracking, occasionally one of the formants was incorrectly categorized as another (e.g., F2 being misinterpreted as F1, etc.). These incorrectly categorized estimates were found and corrected by examining a plot with all of the steady-state F1, F2, and F3 estimates for each individual.

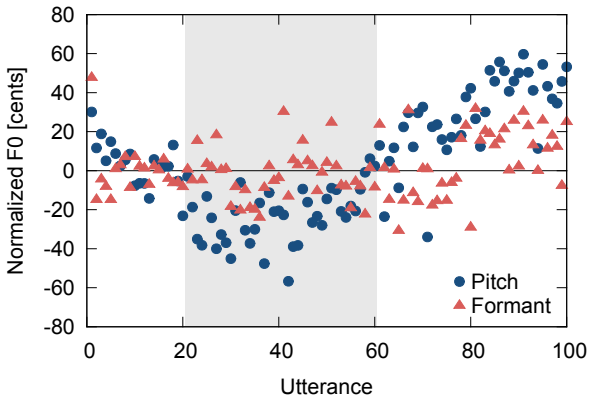


Figure 2: *Average F0 for each utterance in the Pitch (circles) and Formant (triangles) Perturbation conditions. Individuals' F0 results were converted to cents using the average of the last 15 utterances of the Baseline phase as a reference for each individual. The shaded region indicates when the feedback was perturbed.*

3. Results

3.1. Pitch Compensation

Using the F0 results from the Pitch Perturbation condition, a baseline average F0 was calculated for each individual from the last 15 utterances of the Baseline phase (i.e., utterances 6–20). Using their baseline average, individuals' F0 results were then normalized by converting from Hz to cents. The F0 results from the Formant Perturbation condition were analyzed in a similar manner. The normalized results for each utterance, averaged across talkers, can be seen in Figure 2.

From Figure 2, it appears that talkers did not alter F0 during the Formant Perturbation condition, but did alter F0 during the Shift phase of the Pitch Perturbation condition. To quantify this, the compensation of each talker was calculated. Here, the magnitude of compensation was defined as the average normalized F0 (in cents) of the last 15 utterances of the Shift phase. The sign of the compensation was defined as positive if it opposed the perturbation and negative if it followed the perturbation. For the Pitch Perturbation condition, a single sample t-test of talkers' compensations was not significantly different from 0 [$t(21) = 0.945, p > 0.35$]. A closer examination of individual results revealed a wide range of compensation. While 15 talkers compensated (i.e., altered production in the direction opposite the perturbation), 7 of the talkers followed (i.e., altered production in the same direction as the perturbation). Thus, the lack of statistical significance is likely due to the mix of compensators and followers.

3.2. Formant Compensation

Formant compensations were examined in two contexts: the response to a direct formant perturbation (Formant Perturbation session) and the response to an inadvertent shift of the formant when the pitch is shifted with an effects processor (Pitch Perturbation session).

From the formant results from the Formant Perturbation session, a baseline average for F1 and F2 was calculated for each individual from the last 15 utterances of the Baseline phase (i.e., utterances 6–20). Each individual's F1 and F2 results were

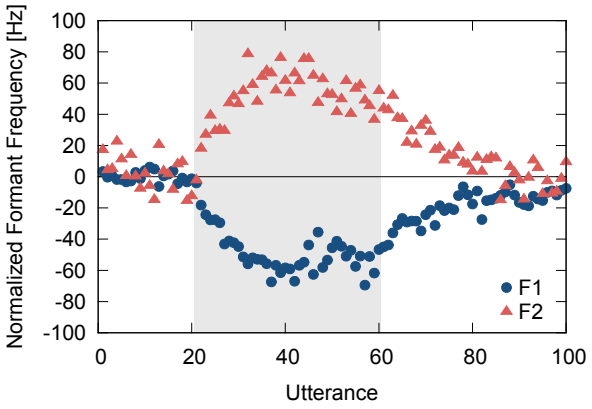


Figure 3: *Average normalized F1 (circles) and F2 (triangles) frequencies for each utterance in the Formant Perturbation condition. The shaded region indicates when the feedback was perturbed.*

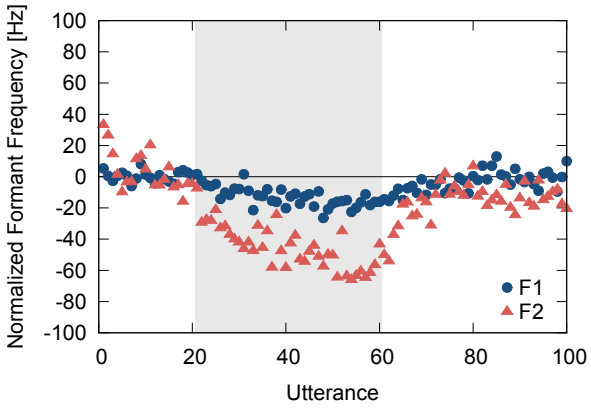


Figure 4: *Average normalized F1 (circles) and F2 (triangles) frequencies for each utterance in the Pitch Perturbation condition. The shaded region indicates when the feedback was perturbed.*

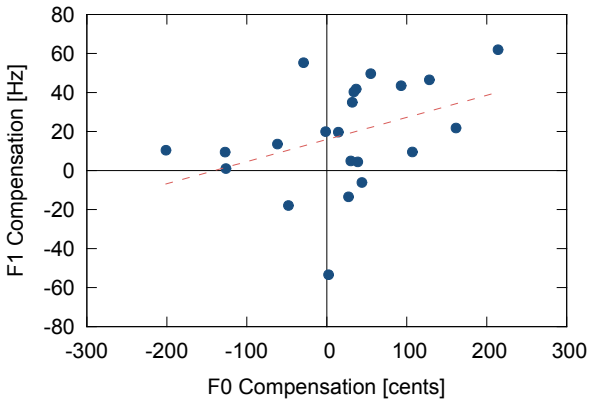


Figure 5: *Scatterplot of talkers' compensation in F0 and F1 in the Pitch Perturbation condition.*

then normalized by subtracting that individual's baseline average. The normalized results for each utterance, averaged across talkers, can be seen in Figure 3.

The pitch shifting algorithm used by the effects processor shifts the entire spectrum. The result is that, along with the pitch, the formant frequencies are also perturbed. Over the last 15 utterances of the Baseline phase of the Pitch Perturbation condition, the average formant frequency produced by talkers was 736.5 and 2048.9 Hz for F1 and F2 respectively. During the Shift phase, the auditory feedback was increased by 200 cents. Thus, on average, the frequencies of F1 and F2 were shifted by 90.2 and 251.0 Hz respectively. For F2, this resulted in a formant shift that was almost identical in magnitude, but opposite in direction, to that applied in the Formant Perturbation condition.

To examine if talkers altered their formants in the Pitch Perturbation condition, a similar normalization process was conducted on the formant results. Again, a baseline average for F1 and F2 was calculated for each individual from the last 15 utterances of the Baseline phase and used to normalize each individual's F1 and F2 results. The normalized results for each utterance, averaged across talkers, can be seen in Figure 4.

From Figures 3 and 4, it is clear that, on average, talkers altered formant production during the Shift phase of both perturbation conditions. To confirm this, repeated measures ANOVAs were conducted with phase (the average formant frequency of the last 15 utterances in the Baseline vs. Shift) as within- and order of the perturbation conditions as between-subject factors. For the results from the Formant Perturbation condition, a significant effect of phase was found for both F1 [$F(1, 20) = 35.566$, $p < 0.001$] and F2 [$F(1, 20) = 15.67$, $p = 0.001$] but no significant effect of order was found for either F1 [$F(1, 20) = 3.77$, $p = 0.07$] or F2 [$F(1, 20) = 1.624$, $p = 0.22$]. Similarly, for the results from the Pitch Perturbation condition, a significant effect of phase was found for both F1 [$F(1, 20) = 9.643$, $p = 0.006$] and F2 [$F(1, 20) = 26.881$, $p < 0.001$] but no significant effect of order was found for either F1 [$F(1, 20) = 2.233$, $p = 0.15$] or F2 [$F(1, 20) = 1.508$, $p = 0.23$].

For each perturbation condition, the compensation in F1 and F2 of each talker was calculated. Here, the magnitude of compensation was defined as the difference between the average formant frequencies of last 15 utterances of the Baseline and Shift phases. Again, the sign of the compensation was defined as positive if it opposed the perturbation and negative if it followed the perturbation. While most compensated, a few talkers followed rather than opposed the formant perturbation; one talker followed in F1, and two in F2, but no talker followed in both F1 and F2.

3.3. Comparison of Pitch and Formant Compensations

Talkers' compensations to formant and pitch perturbations in the different conditions were compared and correlations were computed. A modest correlation was observed between F0 and F1 compensations within the Pitch Perturbation condition [$r(22) = 0.392$, $p = 0.04$, one-tailed; see Fig. 5] but not between F0 compensation in the Perturbation condition and F1 compensation in the Formant Perturbation condition [$r(22) = -0.124$, $p > 0.29$, one-tailed].

A trend for a modest correlation was observed between F1 compensations between the Pitch and Formant Perturbation conditions [$r(22) = 0.300$, $p = 0.09$, one-tailed] but not for F2 compensations between the Pitch and Formant Perturbation conditions [$r(22) = -0.230$, $p = 0.15$, one-tailed].

Further, no significant correlation was observed between compensations in F1 and F2 within the Formant Perturbation condition [$r(22) = 0.159$, $p = 0.24$, one-tailed] or between F1 and F2 within the Pitch Perturbation condition [$r(22) = 0.082$, $p = 0.36$, one-tailed].

4. Discussion

In this study, talkers repeatedly produced utterances of the word "head" while receiving auditory feedback in which the pitch or formant frequencies had been perturbed in real-time. As in previous studies, partial compensations to both pitch and formant perturbations were observed and the magnitude of compensation varied widely across talkers.

One potential explanation for the large variability in compensations observed across talkers may be individual differences in the relative weighting of auditory vs. somatosensory feedback in speech-motor control. While no significant correlation was found between pitch compensations in the Pitch Perturbation condition and formant compensations in the Formant Perturbation condition, a modest correlation was observed between compensations in pitch and formant frequencies within the Pitch Perturbation condition. The observation of a modest correlation supports this relative weighting hypothesis. However, the lack of significant correlation across perturbation conditions suggests that other sources of variability are also involved.

The paradigm used in this experiment is different from that used in a typical pitch perturbation study. In the present study, pitch shifts were applied to an entire utterance rather than being restricted to a short interval midway through the vowel. This provides two advantages. First, it allowed talkers to speak normally rather than prolonging their vowels. Second, it allowed us to examine pitch and formant compensations within the same set of utterances.

The pitch shifting algorithm employed by the effects processor used in this study shifted the entire spectrum of the input signal. Thus, both pitch and formant frequencies were affected. The magnitude of the pitch perturbation used in the present study was 200 cents. This value was chosen because, for the word and talkers used in the study, F2 would be shifted by similar amounts in both the Pitch and Formant Perturbation conditions. On average, in both conditions, the talkers compensated equally. However, individual compensations in F2 were not correlated across conditions. While the formant shift of F1 was smaller in the Pitch Perturbation condition, talkers still compensated, and a trend for modest correlation between individuals' F1 compensations in the Pitch and Formant Perturbation conditions was observed. Thus, while the order of perturbation conditions was not found to have an effect on overall compensation, individuals' responses to formant perturbations varied between conditions. This variability may suggest that the compensatory response may not be as stable as previously thought.

The 200 cent pitch perturbation used in the present study is, in general, larger than most studies of pitch perturbation. Previous perturbation studies have observed that the percentage of compensation (i.e., the compensation divided by the magnitude of the perturbation) decreases for large perturbations [10]. Similarly, formant compensation has been found to be approximately linear for small perturbations, but non-linear, and proportionally smaller, for large perturbations [8, 11]. Thus, the magnitudes of the perturbations used in the present experiment may have resulted in a more linear response to the formant than the pitch perturbation.

Studies that have more closely examined the time course of adaptations to pitch perturbations have identified both volitional and reflexive components of the reponse [12]. In the present study, the measurement of pitch compensation did not explore the effects of these individual components. In contrast, a voluntary component of the response to formant perturbation has not been found [4]. Thus, there are some differences in mechanisms used for speech-motor control of pitch and formants. Future studies that can isolate the components of the pitch response and compare them to the formant response may better test the relative-weighting hypothesis.

Identifying the source of individual differences in compensatory responses remains a difficult task. The results of the present study suggest that the comparison of individual compensations to perturbations of different acoustical characteristics of speech is a promising method to explore this problem.

5. Acknowledgements

This research was supported by the (US) National Institute of Deafness and Communicative Disorders Grant DC-08092

6. References

[1] Kawahara H, "Hearing voice: Transformed auditory feedback effects on voice pitch control," in *Proceedings of the International Joint Conference on Artificial Intelligence: Workshop on Computational Auditory Scene Analysis*, Montreal, Canada, 1995, pp. 143–148.

[2] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation in speech production." *Science*, vol. 279, no. 5354, pp. 1213–1216, 1998.

[3] T. A. Burnett, M. B. Freedland, C. R. Larson, and T. C. Hain, "Voice F0 responses to manipulations in pitch feedback," *Journal of the Acoustical Society of America*, vol. 103, no. 6, pp. 3153–3161, 1998.

[4] K. G. Munhall, E. N. MacDonald, S. K. Byrne, and I. Johnsrude, "Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate." *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 384–90, Jan. 2009.

[5] E. N. MacDonald, D. W. Purcell, and K. G. Munhall, "Probing the independence of formant control using altered auditory feedback." *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 955–65, Feb. 2011.

[6] S. Tremblay, D. M. Shiller, and D. J. Ostry, "Somatosensory basis of speech production." *Nature*, vol. 423, no. 6942, pp. 866–869, 2003.

[7] C. R. Larson, K. W. Altman, H. Liu, and T. C. Hain, "Interactions between auditory and somatosensory feedback for voice F0 control." *Experimental Brain Research*, vol. 187, no. 4, pp. 613–621, 2008.

[8] E. N. MacDonald, R. Goldberg, and K. G. Munhall, "Compensations in response to real-time formant perturbations of different magnitudes." *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 1059–68, Feb. 2010.

[9] IPA, *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press, 1999.

[10] H. Liu and C. R. Larson, "Effects of perturbation magnitude and voice F0 level on the pitch-shift reflex." *Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3671–3677, 2007.

[11] S. Katseff, J. Houde, and K. Johnson, "Partial Compensation for Altered Auditory Feedback: A Tradeoff with Somatosensory Feedback?" *Language and Speech*, in Press. [Online]. Available: http://dx.doi.org/10.1177/0023830911417802

[12] T. A. Burnett, K. E. McCurdy, and J. C. Bright, "Reflexive and volitional voice fundamental frequency responses to an anticipated feedback pitch error." *Experimental Brain Research*, vol. 191, no. 3, pp. 341–51, Nov. 2008.

Prosodic Characteristics of Feedback Expressions in Distracted and Non-distracted Listeners

Zofia Malisz[*], *Marcin Włodarczak*[*], *Hendrik Buschmeier*[†], *Stefan Kopp*[†], *Petra Wagner*[*]

[*]Faculty of Linguistics and Literary Studies
[†]Sociable Agents Group, CITEC and Faculty of Technology
Bielefeld University, Bielefeld, Germany
{zofia.malisz,petra.wagner,mwlodarczak}@uni-bielefeld.de
{hbuschme,skopp}@techfak.uni-bielefeld.de

Abstract

In a previous study [1] we investigated properties of communicative feedback produced by attentive and non-attentive listeners in dialogue. Distracted listeners were found to produce less feedback communicating understanding. Here, we assess the role of prosody in differentiating between feedback functions. We find significant differences across all studied prosodic dimensions as well as influences of lexical form and phonetic structure on feedback function categorisation. We also show that differences in prosodic features between attentiveness states exist, e.g., in overall intensity.

Index Terms: communicative feedback; prosody; dialogue; distraction; engagement; attention

1. Introduction

In spoken dialogue the behaviour of the interlocutor who is currently listening is characterised by short feedback signals (e.g., "uh-huh", "m", "yeah", "okay"). These signals minimally communicate presence, perception, understanding, acceptance as well as higher feedback functions such as agreement and attitudinal reactions to the speaker [2]. Feedback signals play an important role in grounding and coordination of the interaction as they allow listeners to inform speakers of their state of perception, understanding, etc. without interrupting the ongoing turn. At the same time speakers can estimate online how successful their utterance has been in communicating the intended message via received feedback. Therefore, feedback is used to adapt communicative behaviour to the listener's needs.

As a result, communication becomes difficult when feedback is inappropriately timed or expressed. In [3], listeners were induced to produce less context-specific feedback, which had a substantial influence on the speakers' behaviour and the quality of their storytelling. Similarly, in [4], speakers told more vivid stories when they expected an attentive listener and in fact interacted with one. Speakers also spent more time telling their stories when their expectations of listeners' attention states matched reality. Both studies showed that distractedness in listeners had an influence on speakers and their behaviour.

Conversational situations exist, where listeners are being distracted by simultaneous tasks (browsing the Internet, reading documents, etc.) or disengaged for other reasons. Speakers are then mostly able to notice that their dialogue partners are distracted and change their communicative behaviour accordingly. As [4, p. 582] note, speakers are "painfully aware when their conversational partners [. . .] are inattentive, and they can often tell when their partners are only pretending to pay attention." Consequently, it is reasonable to assume that distractedness manifests in the listeners' communicative behaviour in general and their feedback behaviour in particular. Dialogue partners should be able to perceive if listeners' behaviour deviates from the one expected of a fully engaged interlocutor.

Important engagement indicators are the timeliness and frequency of feedback signals in response to feedback elicitation cues produced by the speaker (e.g., [5]). In a multimodal context, listeners display mutual, joint and shared attention with gaze [6]. General presence, liveliness and readiness to cooperate is also signalled with posture shifts, appropriate head movements and manual gestures [7, 8].

The influence of prosody on the pragmatic function of feedback utterances has been a subject of study for some time. Syllabification, duration, loudness, pitch slope and pitch contour were identified as relevant for the discrimination of functional feedback categories in English [9]. A more detailed analysis in [10] found that English affirmative cue words are higher in pitch, intensity and pitch slope when used as a backchannel. Backchannels were also longer in duration, produced with shorter latencies and often preceded by a pitch rise in the interlocutor's speech. A study of Japanese backchannels, however, found that prosodic features marking interest and surprise vary depending on the backchannel's lexical realisation [11]. For German, a cluster analysis of the relation between the linguistic function and intonational form of the discourse particle "hm"revealed prototypical and functionally equivalent variants [12]. [13] used synthesised instances of the German backchannel "ja" with durational features and F_0 curves modelled after [14]. Subjects were asked to evaluate the backchannels along seven semantic dimensions (e.g.: *happy* vs. *sad*). The analysis identified prosodic features related to agreement, happiness, boredom, etc.

In previous work [1], we studied the distribution of functional feedback categories between distracted and attentive listeners in a dialogue corpus collected on the basis of the paradigm in [3]. We found that distracted listeners produced less feedback communicating understanding than attentive listeners. In the present paper, we analyse prosodic characteristics of the three most frequent feedback expressions in our corpus ("ja", "m", and "mhm") across their pragmatic functions. We also examine the differences in feedback produced by distracted vs. attentive listeners.

The first three authors contributed to the paper equally.

2. Data collection

To gather reliable data on feedback behaviour of distracted and attentive listeners, we carried out a lab-based face-to-face dialogue study. One of the dialogue partners (the 'storyteller') told two holiday stories to the other participant (the 'listener'), who was instructed to listen actively, make remarks and ask questions.

Listeners were engaged in a distraction task during either the first or the second story. Building upon the paradigm of [3], we instructed listeners to press a button on a hidden remote control every time the dialogue partner produced a word starting with the letter 's' (the second most common German word-initial letter usually corresponding to perceptually salient sibilants). In addition, they had to count the total number of 's-words'. Storytellers were informed that their partners would be listening for something in the dialogue, but they did not know during which of the two stories.

Participants were seated approximately three metres apart to minimise crosstalk. Close talking high-quality headset microphones were used. Furthermore, another microphone captured the whole scene and a fourth audio channel was used to record the 'clicks' synthesised by a computer when listeners pressed the button on the remote control. Interactions were recorded from three camera perspectives: medium shots showing the storyteller and the listener and a long shot showing the whole scene.

A total of fifty students (34 female and 16 male native speakers of German) were recruited at Bielefeld University to participate in the study, receiving either course credit or 4 euro as payment. They were assigned to one of 25 same-sex dyads. In all but four participant pairs, dialogue partners were unacquainted.

3. Annotation

Many annotation schemes distinguish only between two broad feedback function categories such as 'generic' vs. 'specific' [3].

For our more detailed analysis of listeners' behaviour, an annotation scheme distinguishing between subtler pragmatic variants of feedback signals was needed. The annotation scheme, discussed in detail in [1], is based largely on the framework of [2, 15] ascribing up to four basic functions to feedback signals: *contact, perception, understanding* and *attitudinal reactions*. In communicating one of these functions, listeners express their willingness and ability to continue the interaction, perceive or understand the message or to respond to it. In the present work, we focus on the three affirmative functions, named P1, P2 and P3, where P1 can be seen as what is usually called a backchannel or a 'continuer'. Category P2 signals successful interpretation of the message, and category P3 indicates acceptance, belief and agreement. These levels can be treated as a hierarchy with increasing value of judgement and 'cognitive involvement' or 'depth' of grounding. See Table 1 for an overview.

Feedback utterances in 14 sessions (i.e., 28 dialogues) were segmented and transcribed according to German orthographic conventions (where existent). A total of 1003 feedback functions were annotated, each independently by three annotators taking communicative context into account. Majority labels between annotators were calculated automatically and problematic cases (110; roughly 10%) were discussed and resolved.

4. Feature extraction and analysis

The 28 annotated dialogues have a total length of 180 minutes and each dialogue has a mean length of 6:25 minutes (Min = 2:16; Max = 14:29; SD = 2:31). On average 36 feedback signals were produced per dialogue (Min = 7; Max = 93; SD = 23.1).

Table 1: A subset of the feedback functions inventory. A detailed description can be found in [1].

C	Definition of category
P1	The partner signals perception of the signal. *'I hear you and please continue.'*
P2	The partner signals perception and understanding of the message content. *'I understand what you mean.'*
P3	The partner signals perception, understanding and acceptance of the message or agreement with the message. *'I accept / agree / believe what you say.'*

Duration in milliseconds was calculated for each feedback signal. Pitch and intensity values were extracted using Praat[1]. In order to avoid tracking errors, pitch was extracted in two steps with the floor and ceiling values for the second run set at the 15th percentile times 0.83 and the 65th percentile times 1.92 of the values in the initial run [16]. All measurements were converted to z-scores to normalise the differences between dialogues.

We calculated mean, standard deviation, and slope of pitch and intensity in each feedback signal. Next, we split each feedback signal into three parts of equal length and calculated the mean and standard deviation for each of these parts. Similarly, slopes (from linear regression) were calculated over the first and second half. The procedure yields the following features for each feedback signal: i) dialogue act label, ii) orthographic transcription, iii) duration, iv) mean.[pitch, intensity], v) sd.[pitch, intensity], vi) slope.[pitch, intensity], vii) segment.[1,2,3].mean.[pitch, intensity], viii) segment.[1,2,3].sd.[pitch, intensity], ix) segment.[1,2].slope.[pitch, intensity].

Two separate analyses using Generalised Linear Mixed Models (GLMM) were conducted for (a) feedback function differences and (b) distractedness-related differences. A dataset was used with expressions "ja", "mhm" and "m" combined. The prosodic feature vector exhibited high collinearity even after centring and scaling of the variables. Since high correlations between variables influence the validity of regression estimates for individual predictors, we performed a Principal Component Analysis to deal with multicollinearity. The procedure reduced the feature vector from 23 to 9 dimensions, which were chosen according to the cumulative level of variance explained by the components, here set at 0.94. The Varimax-rotated components were entered into the GLMMs.

A GLMM (with cumulative logit link function) was fitted with Feedback Function (P1, P2 and P3) as a dependent ordinal multinomial variable[2]. The variable Feedback Expression was entered as a fixed factor and in an interaction term with all prosody-based components to account for variability that is due to the phonetic structure of the different expressions. Other fixed factors included Task Order and Experimental Condition. The only random effect entered was Session, equivalent to speaker differences in laboratory designs.

A second GLMM (with logit link function) was fitted with Experimental Condition (distracted vs. non-distracted) as a binomial dependent variable[3]. All other terms were specified as in the model above with the exception of Feedback Function included here as a fixed factor.

[1] `http://www.fon.hum.uva.nl/praat/`

[2] Using the GENLINMIXED command in IBM SPSS Statistics 20.0.

[3] Using the lme4 R package, version 0.999375-42.

Table 2: The first nine components (accounting for 92% of the variance in the dataset; ordered by standardised loadings and proportional variances) alongside the prosodic features with high loadings on each component.

RC_i	Load	Var	Prosodic feature	Load
RC_1	3.60	0.16	segmented.slope.pitch.2	0.92
			segmented.sd.pitch.3	0.84
			segmented.mean.pitch.3	0.75
			slope.pitch	0.75
			sd.pitch	0.71
RC_2	3.24	0.14	mean.intensity	0.98
			segmented.mean.intensity.1	0.86
			segmented.mean.intensity.2	0.85
			segmented.mean.intensity.3	0.77
RC_7	2.90	0.13	segmented.mean.pitch	0.92
			mean.pitch	0.89
			segmented.mean.pitch.1	0.83
RC_3	2.44	0.10	sd.intensity	0.94
			segmented.sd.intensity.3	0.81
RC_4	2.28	0.10	segmented.slope.pitch.1	−0.90
			segmented.sd.pitch.1	0.87
RC_5	2.14	0.09	slope.intensity	0.92
RC_8	1.92	0.08	segmented.slope.intensity.1	−0.87
			segmented.sd.intensity.2	0.70
RC_9	1.35	0.06	segmented.sd.pitch.2	0.84
RC_6	1.23	0.04	duration	0.92

The statistically significant components resulting from each model were interpreted in terms of prosodic features with high loadings on a component (tabulated in Table 2). Thresholds for choosing a feature as relevant were set at the point of clear discontinuity within each component. Notably, all components are interpretable in terms of disjoint and coherent feature sets.

5. Results

5.1. Differences in prosody between feedback functions

Table 3 presents the main effects and interactions found to significantly differentiate between feedback functions in the GLMM described in Section 4.

Following the proportional odds assumption of multinomial ordinal logistic regression, we can interpret the proportional odds of choosing lower categories against higher categories given any partitioning category, i.e., P1 versus P2 and P3 combined, and P1 and P2 combined versus P3. Consequently, as RC_1 increases by one unit, the odds of choosing lower categories increases by 1.325. By contrast, a one-unit change in components RC_2, RC_9, RC_6 decreases the odds of choosing lower categories by 0.793, 0.685 and 0.770 respectively. Moving from "mhm" to "ja" decreases the odds of choosing lower categories. For RC_4 the odds of choosing lower categories depend on the lexical form and decrease by 0.563 for "ja" and increase by 2.002 for "mhm."

Interpreting components in term of prosodic features (Table 2), RC_1 expresses pitch variability especially in the last part of the expression (recall that slope values were calculated for the expression cut in two segments; the other values, SD and mean were calculated for the expression cut in three segments). RC_2 expresses mean intensity. The next component which had a significant main effect on prosodic categorisation of Feedback Function was RC_9, reflecting the variability of pitch in the middle of the expression. RC_6 is essentially duration. RC_4, i.e., pitch variability and the magnitude of the variability (slope) in the first part of the expression interact with the phonetic form of the expression itself. In other words, the effect of RC_4 depends on the particular expression e.g. "ja" and "mhm" or "m". At the same time there is a main effect of the feedback expression form on the feedback function classification, where "ja" highly significantly distinguishes between the categories.

Table 3: Fixed coefficients of the multinomial GLMM for Feedback Function (reference category: P1). LO: log odds; PO: proportional odds; SE: standard error; significance codes: 0.05: $*$, 0.01: $**$.

Model term	LO	PO	SE	t	p
RC_1 $*$	0.281	1.325	0.122	2.306	0.022
RC_2 $*$	−0.232	0.793	0.101	−2.305	0.022
RC_9 $*$	−0.378	0.685	0.117	−3.224	0.001
RC_6 $*$	−0.262	0.770	0.109	−2.404	0.017
"ja" $*$	−1.799	0.165	0.333	−5.398	0.000
"m"	−0.130	0.878	0.284	−0.458	0.647
RC_4×"ja" $*$	−0.574	0.563	0.223	−2.573	0.010
RC_4×"m"	−0.050	0.951	0.146	−0.342	0.733
RC_4×"mhm" $**$	0.694	2.002	0.264	2.633	0.009

5.2. Differences in prosody between conditions

Table 4 presents model estimates of predictors found to differentiate between feedback signals produced by listeners in the distracted and non-distracted condition. The value of the C index measures the concordance between the predicted probability in the model and the observed response. From a value of $C = 0.8$ a model exhibits real predictive power, given the subtle phenomena we are dealing with here, our value of $C = 0.77$ is strong.

The estimate for Feedback Function as a predictor of attentiveness confirms our previous results [1], where a decrease in the frequency of signalling understanding (function P2) was found in distracted listeners. The present model predicts that a unit increase in the P2 function increases the odds of the listener being attentive by 1.716.

As far as prosodic correlates of attentiveness are concerned, the results show that RC_2 (defined by the overall mean intensity measures) is highly significant; attentive speakers tend to speak more loudly. Energy is also less variable in the non-distracted case (RC_3 estimate). RC_1 defined by pitch variability measures is positively related to attentiveness (one-unit increase in RC_1 increases the odds of an attentive state by 1.823). It can be expected that feedback delivered with less intonational variability is indicative of lower engagement.

Significant interactions between the phonetic form of the expression and some of the pitch and intensity measures on the one hand allow for the fine-tuning of potential recognition of attention states in particular expressions and on the other hand confirm the differences depending on phonetic structure.

6. Discussion

The results presented in the previous section indicate that prosody plays a role in distinguishing between different functions of communicative feedback. However, the significant interactions with the lexical form in our results suggests that it is import-

Table 4: Fixed coefficients of the binomial GLMM for Condition. LO: log odds; PO: proportional odds; SE: standard error; significance codes: 0.05: $*$, 0.01: $**$, 0.001: $***$. Predictive strength measures: C = 0.77, $D_{xy} = 0.55$.

Model term	LO	EO	SE	z	p
"m"	0.451	1.570	0.365	1.237	0.216
"mhm"	−0.341	0.711	0.396	−0.862	0.389
RC_1*	0.600	1.823	0.284	2.113	0.035
RC_3*	−0.565	0.568	0.244	−2.319	0.020
RC_4**	−0.853	0.426	0.319	−2.672	0.007
RC_2***	0.472	1.603	0.114	4.144	0.000
RC_6*	−0.260	0.772	0.119	−2.181	0.029
$P2*$	0.540	1.716	0.255	2.120	0.034
P3	0.161	1.175	0.381	0.422	0.673
order$*$	1.109	3.031	0.442	2.507	0.012
$RC_1\times$"m"$**$	−0.965	0.381	0.332	−2.902	0.004
$RC_1\times$"mhm"	−0.517	0.596	0.347	−1.491	0.136
$RC_3\times$"m"	0.437	1.548	0.303	1.441	0.149
$RC_3\times$"mhm"$*$	0.774	2.169	0.316	2.448	0.014
$RC_4\times$"m"$*$	0.838	2.312	0.354	2.365	0.018
$RC_4\times$"mhm"$*$	0.861	2.365	0.403	2.137	0.033

ant to take the phonetic structure of particular expressions into account: prosodic features may strongly depend on segmental structure, e.g.: nasality vs. orality in "m" vs. "ja" and syllabic structure in "mhm" vs. monosyllabic expressions such as "ja". Consequently, in addition to general prosodic features distinguishing between feedback functions there are strategies specific to both the feedback function and the particular expression.

Concerning the prosodic characteristics of distracted listeners that could be detected instrumentally and, possibly, also by the interlocutors, some features that might help detect the listener attention state were found. Additionally, the frequency of signalling understanding [1] remains a consistent non-prosodic cue to distractedness in the listener.

7. Conclusions and future work

This work reported on the prosody of German feedback expressions "ja", "mhm" and "m" in a dialogue corpus where listeners' attention was manipulated by an ancillary task. In this study we have taken first steps towards prototypical prosodic profiles of German feedback functions and diverse predictors of attention. However, the complex interaction between prosody, feedback expressions and pragmatic functions needs to be disentangled possibly using more data and automatic classification techniques.

Findings like these could inform automatic methods for detecting listener's attentive states and the communicative intentions they convey via feedback. This information could be used by artificial conversational agents such as spoken dialogue systems to adapt their communicative behaviour to the needs and expectations of the user.

Acknowledgements – This research is supported by the Deutsche Forschungsgemeinschaft (DFG) at the Center of Excellence in 'Cognitive Interaction Technology' (CITEC) as well as at the Collaborative Research Center 673 'Alignment in Communication'. We thank anonymous reviewers for their helpful comments.

8. References

[1] H. Buschmeier, Z. Malisz, M. Włodarczak, S. Kopp, and P. Wagner, "'Are you sure you're paying attention?' – 'Uh-huh'. Communicating understanding as a marker of attentiveness," in *Proceedings of INTERSPEECH 2011*, Florence, Italy, 2011, pp. 2057–2060.

[2] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, pp. 1–26, 1992.

[3] J. B. Bavelas, L. Coates, and T. Johnson, "Listeners as co-narrators," *Journal of Personality and Social Psychology*, vol. 79, pp. 941–952, 2000.

[4] A. K. Kuhlen and S. E. Brennan, "Anticipating distracted addressees: How speakers' expectations and addressees' feedback influence storytelling," *Discourse Processes*, vol. 47, pp. 567–587, 2010.

[5] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[6] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi, "A model of attention and interest using gaze behavior," in *Proceedings of the 5th International Working Conference on Intelligent Virtual Agents*, Kos, Greece, 2005, pp. 229–240.

[7] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, pp. 140–164, 2005.

[8] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue*, London, UK, 2009, pp. 225–234.

[9] N. Ward, "Pragmatic functions of prosodic features in non-lexical utterances," in *Proceedings of Speech Prosody 2004*, Nara, Japan, 2004, pp. 325–328.

[10] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 2007, pp. 1065–1068.

[11] T. Kawahara, Z.-Q. Chang, and K. Takanashi, "Analysis of prosodic features of Japanese reactive tokens in poster conversations," in *Speech Prosody 2010*, Chicago, IL, 2010, pp. 1–4.

[12] J. E. Schmidt, "Bausteine der Intonation?" in *Neue Wege der Intonationsforschung*, ser. Germanistische Linguistik, J. E. Schmidt, Ed. Hildesheim, Germany: Georg Olms Verlag, 2001, vol. 157-158, pp. 9–32.

[13] T. Stocksmeier, S. Kopp, and D. Gibbon, "Synthesis of prosodic attitudinal variants in German backchannel "ja"," in *Proceedings of Interspeech 2007*, Antwerp, Belgium, 2007, pp. 1290–1293.

[14] K. Ehlich, *Interjektionen*. Tübingen, Germany: Max Niemeyer Verlag, 1986.

[15] S. Kopp, J. Allwood, K. Grammar, E. Ahlsén, and T. Stocksmeier, "Modeling embodied feedback with virtual humans," in *Modeling Communication with Robots and Virtual Humans*, I. Wachsmuth and G. Knoblich, Eds. Berlin: Springer-Verlag, 2008, pp. 18–37.

[16] C. De Looze and S. Rauzy, "Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration," in *Proceedings of INTERSPEECH 2009*, Brighton, UK, 2009, pp. 2919–2922.

Formant compensation responses to altered auditory feedback in English and Vietnamese talkers

Linh L.T. Nguyen, David W. Purcell

Health and Rehabilitation Sciences, Faculty of Health Sciences
National Centre for Audiology
Western University, London, ON, Canada
lnguye9@uwo.ca

Abstract

Previous research has indicated that a person's language history may affect their speech compensation results during altered auditory feedback. The purpose of this experiment was to determine if the interaction of first language and second language vowels will influence compensatory behavior in Vietnamese and English speakers. The experiment included a group of native Vietnamese talkers who learned English as a second language and a group of English monolinguals. F1 discrimination thresholds and vowel goodness ratings for English /ɪ/ as in "hid" on a continuum towards English /ɛ/ as in "head" were determined. Vowel spaces were collected for both languages and auditory feedback of F1 was manipulated during speech production for English and Vietnamese vowels. Results found an asymmetry in compensation between the negative and positive shifts for each vowel. Group differences in speech production were found when the F1 manipulation was dynamic. However, when the F1 manipulation was stable, English /ɪ/ did not elicit group differences, but English /æ/ did. These results suggest some differences in the auditory feedback systems for the two groups.

Index Terms: auditory feedback, language acquisition, speech production

1. Introduction

Auditory feedback, hearing one's voice, may play a role in the detection of speech errors and regulation of speech production. Research has found if this system is perturbed, the speaker will make changes in their speech to correct for the perturbation. This correction by the system has been found to occur unconsciously and automatically [1]. Manipulations of the auditory feedback system have been investigated with voice pitch [2], loudness [3] and spectral characteristics of sounds [4]. Recently, studies have altered vowel formants in real-time to change the vowel quality and measure changes in speech production. These studies have found that speakers will compensate for this formant perturbation in the opposite direction [5, 6, 7]. For instance, when talkers are prompted with the word "head" and their F1 feedback was shifted up towards "had", their production changed in F1 towards "hid".

A cross-language study [8] has found that compensation occurred in other languages such as Japanese and when English is a second language. Mitsuya et al. [8] found that Japanese speakers, who were new to learning English, did not compensate as much as native English speakers when Japanese speakers performed the task with the English vowel /ɛ/ as in "head". They postulated that the use of acoustic feedback may be modulated by language experience and phonemic organizations of the native (L1) and second (L2) languages. In the present study, we examined the role of auditory feedback in native Vietnamese talkers who were experienced speakers of English as a second language.

Theories and studies in language acquisition have shown that L1 and L2 can influence each other resulting in differences in speech production and perception results between monolinguals and bilinguals. These results may have occurred because the L1 and L2 phonemes (vowels and consonants) exist in the same phonological space known as the "common phonological space (CPS) or the "L1-L2 phonetic space" [9]. In the CPS, all the phonemic elements of both languages are present and interact with each other. The Vietnamese bilinguals may have a denser CPS than the English monolinguals. This may affect the talker's sensitivity to speech errors and result in differences in speech compensation during altered auditory feedback. The purpose of the present experiment was to determine if the perceptual organization of L1 and L2 vowels may influence Vietnamese and English speakers' compensatory behavior during speech production when auditory feedback is manipulated.

2. Method

2.1. Subjects

Forty-seven individuals from Western University, city of London, and the Greater Toronto Area were recruited for the current study. The sample was divided into two groups: English monolinguals (n = 21 [Females = 17, Males = 4]) and Vietnamese with English as a second language (n = 21 [Females = 17, Males = 4]). English participants had learned English in Ontario, Saskatchewan and British Columbia. All the Vietnamese participants learned English as a second language in Ontario. For each ear, hearing thresholds were measured at octave intervals between 250 Hz and 4 kHz. All individuals had normal thresholds (≤ 25 dB HL). No participants had known language or speech impairments.

Data from five native English participants were discarded because they either learned English in the Maritimes, were not born in Canada, had a history of a lisp or a stutter.

2.2. Equipment and online formant shifting

Equipment and formant analysis were similar to Purcell and Munhall [6]. Participants wore a headset microphone and after amplification the signal was low-pass filtered at a cut-off frequency of 4500 Hz and digitized at 10 kHz sample rate. During altered auditory feedback, the signal was filtered in real time to create shifts in F1. Headphone speech was approximately 80 dBA SPL with background speech shaped noise of 50 dBA SPL. The voice was filtered in real time to create the formant manipulations using an iterative Burg algorithm [10] to estimate formants. Model order was selected for each talker and vowel to optimize the stability of formant estimates [11], which was generally very good. The formant frequency estimates and corresponding filter coefficients were updated every 900 µs. The delay of the auditory feedback was less than 1 ms, but formant calculations included past samples for an effective delay in formant estimates and corresponding filter coefficients of 10 to 20 ms.

2.3. Offline formant analysis

A semi-automated process was used to trim the utterance before and after the vowel. The boundaries were examined by the experimenter and corrected if required.

Steady-state values of F1, F2, and F3 for each utterance were estimated offline [6]. These single values for each formant for each trial were calculated by averaging formant estimates from the middle 80% of the vowel. Occasionally, formants were incorrectly categorized as another (i.e. F1 being categorized as F2, etc.). These errors were found and corrected by examining a graph with all the F1, F2, and F3 values for each participant.

2.4. Speech perceptual tasks

2.4.1. F1 Discrimination Threshold

The minimum change in F1 that the listener can detect was determined for /ɪ/ using a 2-alternative forced choice test on a continuum between /ɪ/ and /ɛ/. A continuum of "hid" was created by shifting F1 upwards in 5 Hz steps. The unaltered "hid" was produced by a young adult male who spoke Canadian English as his first language.

2.4.2. Vowel Goodness Ratings

Goodness is defined as the ability of an exemplar of a specific sound to fit into its respective category [12]. Vowel goodness ratings were used to determine the goodness of various exemplars of English /ɪ/. Eleven versions of "hid" were created, with one unaltered "hid" and ten altered versions, where F1 of the unaltered "hid" was shifted upwards in 20 Hz steps to +200 Hz (i.e. +20, +40,... +200 Hz). The participant was asked to rate each sound on a scale of 1 to 7, where 1 is a very poor version and 7 is an excellent version of "hid".

2.4. Procedure and experimental conditions

Participants were first asked to complete the vowel goodness ratings and 2-alternative forced choice test. Subsequently, in a sound booth, talkers' vowel spaces were collected. Auditory feedback was then altered with F1 shifted either positively or negatively for the English vowels /ɪ/ and /æ/. Both shift directions were also employed with Vietnamese talkers producing the Vietnamese vowel /ɐ/. Talkers said the word "hid", "had" or "tăm" 114 times for each shift direction, as shown in Figure 1. These utterances were grouped into one of four phases. In the first phase, Acclimatization (first 15 utterances), participants received normal feedback. These utterances were discarded prior to analyses. In the second phase, Baseline (utterances 16-35), participants received normal feedback. In the third phase, Ramp (utterances 36-95), F1 was perturbed with the magnitude of the perturbation increasing by 20 Hz after every sixth utterance, resulting in a 200 Hz shift by utterance 90. Finally, in the Hold phase (utterance 96-114), the F1 perturbation at 200 Hz was held constant. The order of the vowels was English /ɪ/, English /æ/ and if the participant was Vietnamese, /ɐ/ was last. In the English /ɪ/ condition, the F1 positive shift was always presented first, followed by the negative shift. Subsequently, shifts were counterbalanced for English /æ/ and Vietnamese /ɐ/. Participants read English reading passages with their headphones off to normalize their speech productions after each directional shift, as well as before and after vowel space collections. Vietnamese participants read Vietnamese passages during the Vietnamese part of the study.

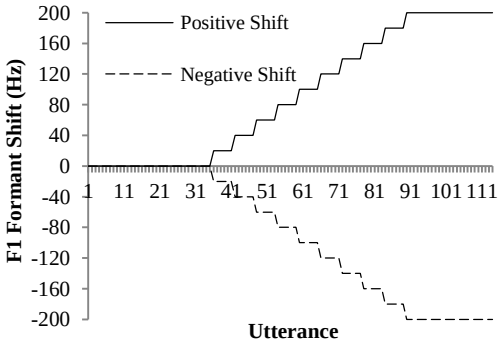


Figure 1: *Schematic procedure of F1 formant shifting.*

3. Results

3.1. Speech perception and vowel space comparisons

Vietnamese speakers' F1/F2 values of the English vowels were compared with those of English monolinguals. The English vowels produced by the female English monolinguals are plotted in Figure 2. The Vietnamese

vowels produced by the female Vietnamese bilinguals are plotted in Figure 3. In these Figures, the center of each ellipse represents the mean F1/F2 frequency for that vowel, while the solid and dashed ellipses represent one and two standard deviations respectively.

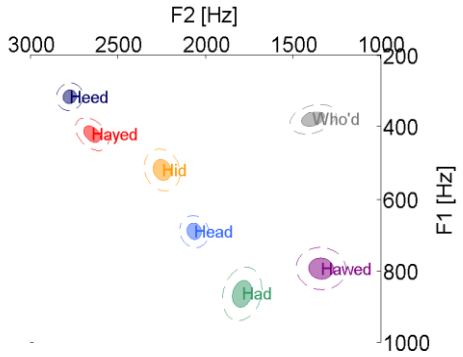


Figure 2: *English vowel space of female, English monolingual speakers in an /hVd/ context (n=17)*

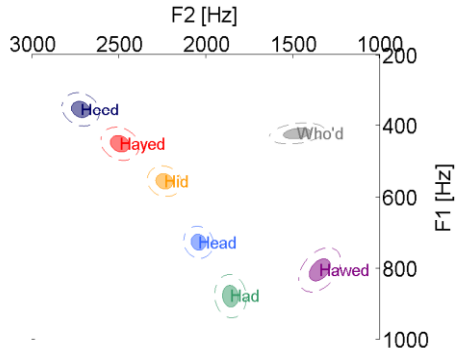


Figure 3: *English vowel space of female, Vietnamese bilingual speakers in an /hVd/ context (n = 17)*

The English vowel spaces of the English and Vietnamese talkers exhibited some differences, but were generally quite similar. The F1discrimination thresholds and Vowel Goodness Ratings for English /ɪ/ did not differ statistically between the groups.

3.2. Compensation comparisons

For each individual, the change in F1 production was normalized by subtracting the average F1 of the Baseline phase from the average F1 of each step (6 utterances per step) in the Ramp phase or the 19 utterances from the Hold phase.

In the following ANOVA analyses, absolute values of the normalized data from each participant were used because of the different signs of the positive and negative shifts. A Sign test was used to evaluate group differences for compensation of English vowels /ɪ/ and /æ/. In the Sign test, the number of mean trials where one group's compensation exceeded the other was calculated from threshold to the end of the Ramp phase and then separately for the Hold phase. Threshold was defined as a change in F1 production two standard deviations from Baseline (refer to ++ symbols in Figures 4 and 5).

The results for the F1-negative condition for the English vowel /ɪ/ are plotted in Figure 4. The results for the F1-positive condition for English /ɪ/ are plotted in Figure 5. Differences in the magnitudes of compensations between the negative and positive shifts were observed. Talkers compensated more in the positive shift than in the negative shift (ANOVA, $F[1,1] = 6.707$, $p < 0.013$). Group differences were observed only in the negative shift during the Ramp phase, where Vietnamese talkers compensated more than English talkers (Sign test, $p < 0.0002$). During the Hold phase, where the F1 shift was held constant at ± 200 Hz, the two groups did not differ in their compensation (positive shift: Sign test ($p = 0.359$); negative shift: Sign test ($p = 0.648$).

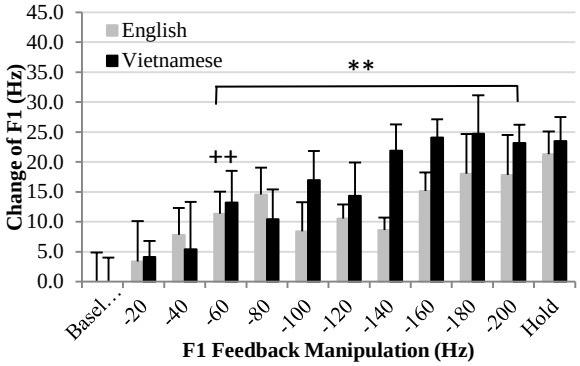


Figure 4: *Average normalized F1 results for the English /ɪ/ as in "hid" across trials for English and Vietnamese speakers during the negative condition. Error bars represent one standard deviation. ++ represents speech compensation threshold. ** represents significant difference between groups.*

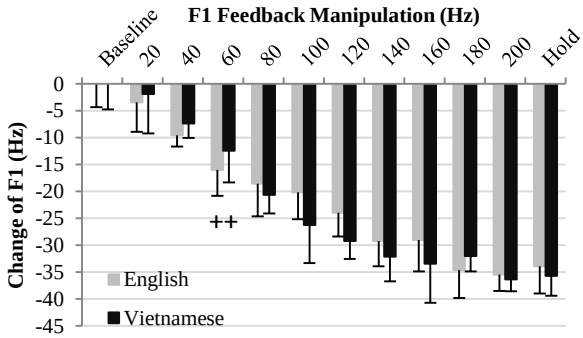


Figure 5: *Average normalized F1 results for the English /ɪ/ as in "hid" across trials for English and Vietnamese speakers during the positive condition. Error bars represent one standard deviation. ++ represents speech compensation threshold.*

Results for English /æ/ are in contrast to English /ɪ/. Differences in the magnitudes of compensations between the negative and positive shifts were also observed. However, talkers compensated more in the negative shift than in the positive shift (ANOVA, $F[1,1] = 5.828$, $p < 0.020$). Group differences were observed in the positive shift

during the Ramp phase (Sign test, $p < 0.0001$) and the Hold phase (Sign test, $p < 0.0044$), where English talkers compensated more than Vietnamese talkers.

The Vietnamese talkers showed compensation when presented with altered auditory feedback for the Vietnamese vowel /ɐ/. An asymmetry of compensation again occurred, where the positive shift had greater compensation than the negative shift (ANOVA, $F[1,1] = 5.683$, $p < 0.027$).

All experimental conditions showed a significant effect for shift size, where an increase of shift size resulted in greater compensation (ANOVA, /ɪ/: $F[1,10] = 27.946$, $p < 0.001$; /æ/: $F[1,10] = 21.780$, $p < 0.001$; /ɐ/: ($F[1,10] = 14.391$, $p < 0.001$).

4. Discussion

The purpose of the current study was to examine the compensatory responses to altered auditory feedback across English and Vietnamese talkers. Both groups received perturbations in feedback where F1 of English vowels /ɪ/ and /æ/ were increased or decreased. Both groups had similar F1 discrimination thresholds for English /ɪ/ and vowel goodness ratings for the different "hid". As well, Vietnamese and English talkers had similar asymmetries in compensation for different shift directions with each vowel. This asymmetry is similar to that reported by Purcell, MacDonald, and Munhall [13]. These results suggest that the Vietnamese talkers produced and perceived English sounds similar to the English monolinguals. This was consistent with the Vietnamese talkers' report that English was presently their dominant language.

However, despite these similarities in the English vowel spaces and discrimination, the dynamic manipulation of F1 in the Ramp phase elicited differences in speech production between the two groups. When F1 was decreased for English /ɪ/, Vietnamese talkers compensated more than English talkers. As well, when F1 was increased for English /æ/, the English talkers compensated more than the Vietnamese talkers. Even though English is presently the Vietnamese talkers' dominant language, their sensitivities to rapid changes in English speech sounds and their ability to compensate to these rapid changes may be different compared to English talkers, leading to the observed group differences during the Ramp phase. In contrast, when F1 was stable during the Hold phase at ±200 Hz, there were no group differences, except for the positive shift for English /æ/. This result may suggest an effect of the organizations of the Vietnamese and English vowel spaces. The Vietnamese vowel space is similar to the English vowel space around English /ɪ/. However, there are differences in the Vietnamese and English vowel spaces around English /æ/. Thus, group differences may have been caused by the influence of the Vietnamese talkers' native language, despite reporting English as their dominant language. Further analyses and experiments are needed to fully understand the differences in compensation results between dynamic and stable manipulations of F1 during altered auditory feedback

5. Acknowledgements

This research was funded by the Ontario Early Researcher Award, the (Canadian) Natural Sciences and Engineering Research Council (NSERC), and supported by the Ontario Graduate Scholarship. The authors would like to thank Takashi Mitsuya, Dr. Ewen MacDonald, Dr. Debra Jared, and Dr. Lisa Archibald for their help and support.

6. References

[1] Munhall, K., MacDonald, E., Byrn, S. and Johnsrude, I., "Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate," J. Acoust. Soc. Am., 125(1):384-390, 2009.

[2] Jones, J.A., and Munhall, K.G., "The role of auditory feedback during phonation: studies of Mandarin tone production," J. Phonetics, 30: 303-320, 2002.

[3] Bauer, J.J., Mittal, J., Larson, C.R. and Hain, T.C., "Vocal responses to unanticipated perturbations in voice loudness feedback an automatic mechanism for stabilizing voice amplitude," J. Acoust. Soc. Am., 119:2363-2371, 2006.

[4] Garber, S., Seigel, G. and Pick, H., "Regulation of vocal intensity in the presence of feedback filtering and amplification," J. Speech and Hearing Research, 24:104-108, 1981.

[5] Purcell, D.W. and Munhall, K.G., "Compensation following real-time manipulation of formants in isolated vowels," J. Acoust. Soc. Am., 119 (4):2288-2297, 2006.

[6] Purcell, D.W. and Munhall, K.G., "Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation," J. Acoust. Soc. Am., 120(2): 966-977, 2006.

[7] MacDonald, E., Goldberg, R. and Munhall, K., "Compensations in response to real time formant perturbations of different magnitudes," J. Aocust. Soc. Am., 127(2):1059-1068, 2010.

[8] Mitsuya, T., MacDonald, E.N., Purcell, D.W. and Munhall, K.G., "A cross language study of compensation in response to real-time formant perturbation," J. Acoust. Soc. Am., 130(5):2978-2986, 2011.

[9] Flege, J., "Second-language speech learning: theory, findings, and problems," in W. Strange (Ed), Speech perception and linguistic experience: Theoretical and methodological issues, 223-273, York, 1995.

[10] Orfandidis, S.J., "Optimum signal processing, an introduction," MacMillan, 1988.

[11] Vallabha, GK. And Tuller, B., "Systematic errors in the formant analysis of steady state vowels," *Speech Communication*, 38(1-2): 141-160, 2002.

[12] Kuhl, P., "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not," *Perception & Psychophysics*, 50(2):93-107, 1991.

[13] Purcell, D.W., MacDonald, E. and Munhall, K., "Cross-vowel use of auditory feedback in English," International Seminar on Speech Production, 2011.

Assessing the Intelligibility and Quality of HMM-based Speech Synthesis with a Variable Degree of Articulation

Benjamin Picart, Thomas Drugman, Thierry Dutoit

TCTS Lab, Faculté Polytechnique (FPMs), University of Mons (UMons), Belgium
{benjamin.picart,thomas.drugman,thierry.dutoit}@umons.ac.be

Abstract

This paper focuses on the assessment of both the intelligibility and the quality of speech when using a variable degree of articulation (hypo/hyperarticulation) in the framework of HMM-based speech synthesis. Intelligibility is evaluated when the synthesizer is working in adverse conditions. The adaptation of a neutral speech synthesizer to generate hypo and hyperarticulated speech is first performed. Simulated noisy and reverberant conditions are then applied to the speech produced by the latter synthesizers. The intelligibility of the resulting speech is assessed by a Semantically Unpredictable Sentences (SUS) test. Results of this test quantify how the possibility of varying the degree of articulation improves the intelligibility of synthetic speech in various adverse conditions. In a second test, natural and synthetic speech quality is evaluated through an Absolute Category Rating (ACR) test. This test allows the assessment of hypo/hyperarticulated speech through various dimensions: comprehension, pleasantness, non-monotony, naturalness, fluidity and pronunciation.

Index Terms: Speech Synthesis, HTS, Expressive Speech, Speaking Style Adaptation, Voice Quality, Speech Intelligibility

1. Introduction

The "H and H" theory [1] proposes two degrees of articulation of speech: hyperarticulated speech, for which speech clarity tends to be maximized, and hypoarticulated speech, where the speech signal is produced with minimal efforts. Therefore the degree of articulation provides information on the motivation/personality of the speaker vs. the listeners [2]. Speakers can adopt a speaking style allowing them to be understood more easily in difficult communication situations. The degree of articulation is characterized by modifications of the phonetic context, of the speech rate and of the spectral dynamics (vocal tract rate of change). The common measure of the degree of articulation consists in defining formant targets for each phone, taking coarticulation into account, and studying differences between real observations and targets vs. the speech rate. Since defining formant targets is not an easy task, Beller proposed in [2] a statistical measure of the degree of articulation by studying the joint evolution of the vocalic triangle area and the speech rate.

Hypo/hyperarticulated speech synthesis has many applications: expressive voice conversion (e.g. for embedded systems and video games), "reading speed" control for visually impaired people (i.e. fast speech synthesizers, more easily produced using hypoarticulation), etc.

This paper is in line with our previous works on expressive speech synthesis [3] [4] [5]. We here focus on the synthesis of different speaking styles, with a varying degree of articulation: neutral speech, hypoarticulated (or casual) and hyperarticulated (or clear) speech. "Hyperarticulated speech" refers to the situation of a teacher/speaker talking in front of a large audience (important articulation efforts have to be made to be understood by everybody). "Hypoarticulated speech" refers to the situation of a person talking in a narrow environment or very close to someone (few articulation efforts have to be made to be understood). It is worth noting that these three modes of expressivity are neutral on the emotional point of view, but can vary amongst speakers, as reported in [2]. The influence of emotion on the articulation degree has been studied in [6] [7] and is out of the scope of this work.

In our previous work on the topic [3], an HMM-based speech synthesizer was built for each degree of articulation (neutral, hypo and hyper) using a large database for each degree of articulation. We then studied the efficiency of speaking style adaptation as a function of the size of the adaptation database [4]. Speaker adaptation [8] is a technique to transform a source speaker's voice into a target speaker's voice, by adapting the source HMM-based model (which is trained using the source speech data) with a limited amount of target speech data. The same idea lies for speaking style adaptation [9] [10]. We were therefore able to produce neutral/hypo/hyperarticulated speech directly from the neutral synthesizer. We finally implemented a continuous control (tuner) of the degree of articulation on the neutral synthesizer [4]. This tuner was manually adjustable by the user to obtain not only neutral/hypo/hyperarticulated speech, but also any intermediate, interpolated or extrapolated articulation degrees, in a continuous way. Finally, we conducted a perceptual evaluation in [5] in order to have a deeper understanding of the phenomena responsible in the perception of the degree of articulation. Starting from an existing standard neutral voice with no hypo/hyperarticulated recordings available, the ultimate goal of our research is to allow for a continuous control of its articulation degree.

In this work, we evaluate the necessity of integrating a variable degree of articulation in a HMM-based speech synthesis system when this latter is embedded in adverse conditions. This situation happens very often in concrete applications: for example, GPS voice inside a moving car (additive noise), train/flight information in stations/halls (reverberation), etc. Does hypo/hyperarticulated speech provides better intelligibility performance in adverse environments? What are the advantages brought by hypo/hyperarticulated speech compared to the standard neutral speech? This work is designed to provide an answer to these two questions.

This paper is structured as follows. After a brief description of the contents of our database in Section 2, the implementation of our synthesizers in the HMM-based speech synthesis system HTS [11] is detailed in Section 3. Speech intelligibility in both noisy and reverberant environments, and speech quality evalu-

ations are performed in Sections 4 and 5. These tests quantify the usefulness of integrating the degree of articulation within HMM-based speech synthesis. Finally Section 6 concludes the paper.

2. Database with various Degrees of Articulation

For the purpose of our research, a new French database was recorded in [3] by a professional male speaker, aged 25 and native French (Belgium) speaking. The database contains three separate sets, each set corresponding to one degree of articulation (neutral, hypo and hyperarticulated). For each set, the speaker was asked to pronounce the same 1359 phonetically balanced sentences (around 75, 50 and 100 minutes of neutral, hypo and hyperarticulated speech respectively), as neutrally as possible from the emotional point of view. A headset was provided to the speaker for both hypo and hyperarticulated recordings, in order to induce him to speak naturally while modifying his articulation degree (see [3] for details on how this was induced).

3. Implementation of the Speech Synthesizers

An HMM-based speech synthesizer [12] was built, relying on the implementation of the HTS toolkit (version 2.1) publicly available in [11]. 1220 neutral sentences sampled at 16 kHz were used for the training, leaving around 10% of the database for synthesis. For the filter, we extracted the traditional Mel Generalized Cepstral (MGC) coefficients (with α = 0.42, γ = 0 and order of MGC analysis = 24). For the excitation, we used the Deterministic plus Stochastic Model (DSM) of the residual signal proposed in [13], since it was shown to significantly improve the naturalness of the delivered speech. More precisely, both deterministic and stochastic components of DSM were estimated on the training dataset for each degree of articulation. In this study, we used 75-dimensional MGC parameters (including Δ and Δ^2). Moreover, each covariance matrix of the state output and state duration distributions were diagonal.

For each degree of articulation, this neutral HMM-based speech synthesizer was adapted using the Constrained Maximum Likelihood Linear Regression (CMLLR) transform [14] [15] in the framework of the Hidden Semi Markov Model (HSMM) [16], with hypo/hyperarticulated speech data to produce a hypo/hyperarticulated speech synthesizer.

In the following, the full data models refer to the models trained on the entire training sets (1220 sentences, respectively neutral, hypo and hyperarticulated), and the adapted models are the models adapted from the neutral full data model, using hypo/hyperarticulated speech data. We showed in [4] that good quality adapted models can be obtained when adapting the neutral full data model with around 100-200 hypo/hyperarticulated sentences. On the other hand, the more adaptation sentences, the better the quality independently of the degree of articulation. This is why we chose in this work to adapt the neutral full data model using the entire hypo/hyperarticulated training sets.

Finally, a linear interpolation of the means and the covariance matrices of each state output and state duration probability density functions (mel-cepstrum, log F0 and duration distributions) is computed between the neutral full data model and the adapted models. For experiments in this work, we chose an interpolation ratio equal to 0.5, corresponding to a model right between the neutral full data model and, on the one hand the adapted hypoarticulated model, and on the other hand the adapted hyperarticulated model.

4. Semantically Unpredictable Sentences Test

In order to evaluate the intelligibility of a voice, the Semantically Unpredictable Sentences (SUS) test was performed on speech degraded alternatively by an additive or a convolutive noise. The advantage of such sentences is that they are unpredictable, meaning that the listeners cannot determine a word in the sentence by the meaning of the whole utterance or the context within the sentence.

4.1. Building the SUS Corpus

The same corpus as the one built in [17] was used in our experiments. This corpus is part of the ELRA package (ELRA-E0023). Basically, 288 semantically unpredictable sentences were generated following 4 syntactic structures:

- adverb det. Noun_1 Verb-t-pron. det. Noun_2 Adjective?
- determiner Noun_1 Adjective Verb determiner Noun_2.
- det. Noun_1 Verb_1 determiner Noun_2 qui (that) Verb_2.
- determiner Noun_1 Verb preposition determiner Noun_2.

Structure 3 originally proposed by [18] was not kept, because it only contained 3 target words (nouns, verbs or adjectives, here written with a capital initial letter) instead of 4 in the other structures. For more details about the generation of this corpus, the reader is referred to [17].

4.2. Procedure

Nineteen people, mainly naive listeners, participated to this evaluation. They were asked to listen to 40 SUS, randomly chosen from the SUS corpus built in the previous paragraph. The SUS were played one at a time. For each of them, listeners were asked to write down what they heard. During the test, listeners were allowed to listen to each SUS at most two times. They were of course not allowed to come back to previous sentences after validating their decision.

The SUS were synthesized using the five synthesizers described in Section 3: neutral (0), hypo (-1) and hyperarticulated (1), interpolated between neutral and hypo (-0.5), and interpolated between neutral and hyper (0.5).

For simulating the noisy environment, a car noise was added to the original speech waveform at two Signal-to-Noise Ratios (SNRs): -5dB and -15dB. The car noise signal was taken from the Noisex-92 database [19], and was added so as to control the overall SNR without silence removal. Since the spectral energy of the car noise is mainly concentrated in the low frequencies (<400Hz), the formant structure of speech was only poorly altered, and voices remained somehow understandable even for SNR values as low as -15dB.

When the speech signal $s(n)$ is produced in a reverberant environment, the observation $x(n)$ at the microphone is:

$$x(n) = h(n) * s(n), \quad (1)$$

where $h(n)$ is the L-tap Room Impulse Response (RIR) of the acoustic channel between the source and the microphone. RIRs are characterized by the value T_{60}, defined as the time for the amplitude of the RIR to decay to -60dB of its initial value. In order to produce reverberant speech, a room measuring 3x4x5 m

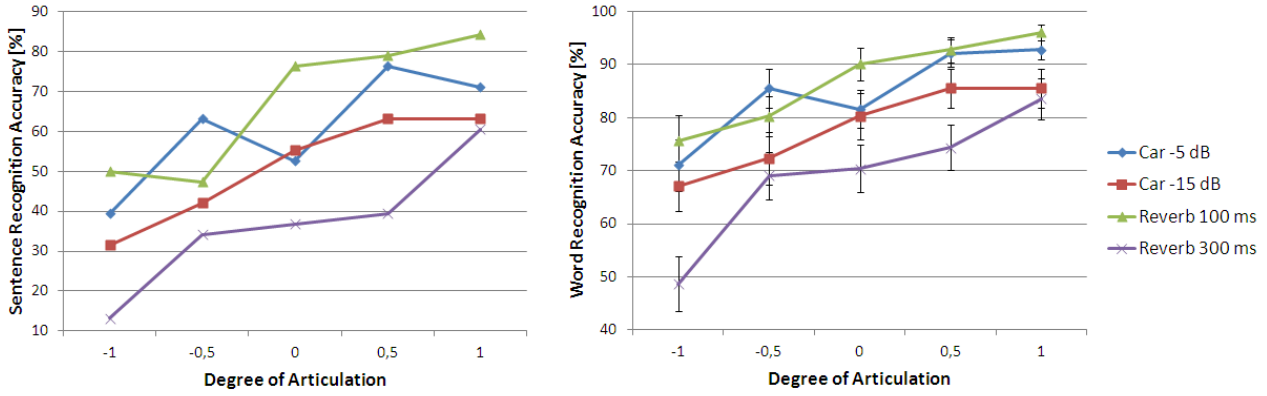


Figure 1: SUS Test - Mean sentence (left) and word (right) recognition accuracies [%].

with two levels of reverberation (T_{60} of 100 and 300ms) was simulated using the source-image method [20], and the simulated impulse responses convolved with original speech signals.

4.3. Results

The mean recognition accuracies at the sentence and word levels (for each degree of articulation, for each type and level of perturbation) are shown in Figure 1. The higher the score, the better the synthesizer intelligibility as it leads to higher sentence/word recognition accuracies. 95% confidence intervals are also displayed for information.

In order to cope with orthographic mistakes, these accuracies were manually annotated, by counting the number of erroneous phonemes for each sentence and word written by the listeners, in comparison with the correct sentence and word. A strong correlation is noted between the recognition accuracy at the sentence and word levels. For the computation of the results, a single erroneous phoneme inside the sentence leads to consider the sentence as wrong. The same idea was applied to the word recognition accuracy computation. Therefore, a sentence could be considered as wrong while some of its words could be considered as correct. Interestingly, accuracy increases with the degree of articulation, and decreases when the perturbation level rises. The worst adverse condition turns out to be the most severe reverberation. Finally, it is noted that the two values of hyperarticulated degree (0.5 and 1) lead to almost the same performance in a car noise, which implies that there is no need to increase the degree of articulation beyond 0.5 in such an environment. This conclusion however does not hold with a reverberant perturbation.

5. Absolute Category Rating Test

Finally, an Absolute Category Rating (ACR) test was conducted in order to assess the quality of speech. As in [17], the Mean Opinion Score (MOS) was complemented with six other categories: comprehension, pleasantness, non-monotony, naturalness, fluidity and pronunciation.

5.1. Procedure

Seventeen people, mainly naive listeners, participated to this evaluation. They were asked to listen to 18 meaningful sentences, randomly chosen amongst the held-out set of the database (used neither for training nor for adaptation). The sentences were played one at a time. For each of them, listeners were asked to rate according to the 7 aspects cited above (for the detailed questions list, see [17]). Listeners were given 7 continuous scales (one for each question to answer) ranging from 1 to 5. These scales were extended one point further on both sides (ranging therefore from 0 to 6) in order to prevent border effects. The sentences corresponded either to the original speech or to the synthesized speech with a variable degree of articulation (neutral, hypo/hyperarticulated). During the test, listeners were allowed to listen to each sentence as many times as wanted. However they were not allowed to come back to previous sentences after validating their decision.

5.2. Results

Results together with their 95% confidence intervals are shown in Figure 2. In all cases, original speech is preferred to synthetic speech. The MOS test shows that original neutral speech is preferred to hypo/hyperarticulated speech, while synthetic neutral and hyperarticulated speech are almost equivalent, leaving synthetic hypoarticulated speech slightly below. The comprehension test points out that neutral and hyperarticulated speech are clearly more understandable than hypoarticulated speech, both on the original and synthetic side. Differences of comprehension between original and synthesized speech are interestingly weak. The pleasantness test indicates a preference of the listeners for original neutral speech, followed by hyper and hypoarticulated speech, while all the types of synthetic speech are equivalently preferred. Despite the HMM modeling, the intonation and dynamics of the voice is well reproduced at synthesis, as illustrated with the non-monotony test. A major problem with HMM-based speech synthesis is the naturalness of the generated speech compared to the original speech. This is a known problem related in many studies. The naturalness test underlines again this conclusion. The fluidity test has an "inverse" tendency compared to other tests. Indeed hypoarticulated speech has a higher score than the others. This is due to the fact that hypoarticulated speech is characterized by a lower number of breaks and glottal stops, shorter phone durations and higher speech rate (as proven in [3]). All these effects lead to an impression of fluidity in speech, while the opposite tendency is observed in hyperarticulated speech. Finally, the pronunciation test correlates with the comprehension test in the sense that the more pronunciation problems are found, the harder the understandability of the message.

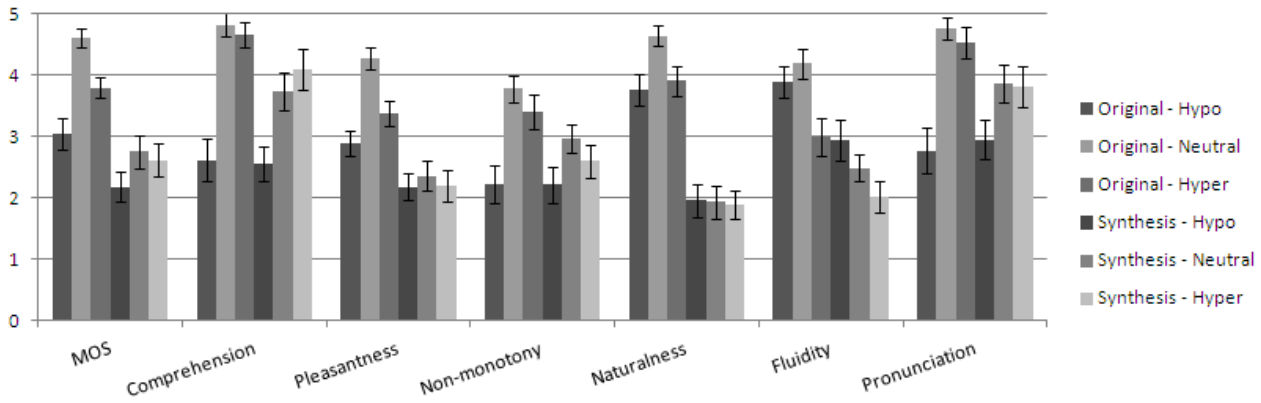


Figure 2: ACR Test - Mean score together with their 95% confidence intervals.

6. Conclusions

This paper focused on the evaluation of HMM-based speech synthesis integrating a variable degree of articulation. In a first step, the intelligibility of speech was assessed through a SUS test. In presence of a perturbation, this evaluation showed that hyperarticulated speech enhances the comprehension of synthetic speech. Moreover, a degree of articulation of 0.5 (instead of 1) is sufficient to improve the recognition of the message in a car noise. The same conclusion is drawn in reverberant environments, except that a degree of articulation of 1 is in this case necessary. In a second step, the quality of speech was assessed by an ACR test. This evaluation showed the gap in naturalness and pleasantness between original speech and synthetic speech. However, it is worth emphasizing that comprehension, non-monotony and pronunciation are well reproduced after the HMM modeling.

Audio examples related to our studies are available online at http://tcts.fpms.ac.be/~picart/.

7. Acknowledgements

Benjamin Picart is supported by the “Fonds pour la formation à la Recherche dans l’Industrie et dans l’Agriculture” (FRIA). Thomas Drugman is supported by the Walloon Region, SPORTIC project #1017095.

8. References

[1] B. Lindblom, Economy of Speech Gestures, vol. The Production of Speech, Spinger-Verlag, New-York, 1983.

[2] G. Beller, *Analyse et Modèle Génératif de l’Expressivité - Application à la Parole et à l’Interprétation Musicale*, PhD Thesis (in French), Universit Paris VI - Pierre et Marie Curie, IRCAM, 2009.

[3] B. Picart, T. Drugman, T. Dutoit, *Analysis and Synthesis of Hypo and Hyperarticulated Speech*, Proc. Speech Synthesis Workshop 7 (SSW7), pp. 270-275, Kyoto, Japan, 2010.

[4] B. Picart, T. Drugman, T. Dutoit, *Continuous Control of the Degree of Articulation in HMM-based Speech Synthesis*, Proc. Interspeech, pp. 1797-1800, Firenze, Italy, 2011.

[5] B. Picart, T. Drugman, T. Dutoit, *Perceptual Effects of the Degree of Articulation in HMM-based Speech Synthesis*, Proc. NOLISP Workshop, pp. 177-182, Las Palmas, Gran Canaria, 2011.

[6] G. Beller, *Influence de l’expressivité sur le degré d’articulation*, RJCP, France, 2007.

[7] G. Beller, N. Obin, X. Rodet, *Articulation Degree as a Prosodic Dimension of Expressive Speech*, Fourth International Conference on Speech Prosody, Campinas, Brazil, 2008.

[8] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, S. Renals, *A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis*, IEEE Audio, Speech, & Language Processing, vol. 17, no. 6, pp. 1208-1230, August 2009.

[9] J. Yamagishi, T. Masuko, T. Kobayashi, *HMM-based expressive speech synthesis – Towards TTS with arbitrary speaking styles and emotions*, Proc. of Special Workshop in Maui (SWIM), 2004.

[10] T. Nose, M. Tachibana, T. Kobayashi, *HMM-Based Style Control for Expressive Speech Synthesis with Arbitrary Speaker’s Voice Using Model Adaptation*, IEICE Transactions on Information and Systems, vol. 92, no. 3, pp. 489-497, 2009.

[11] [Online] HMM-based Speech Synthesis System (HTS) website : http://hts.sp.nitech.ac.jp/

[12] H. Zen, K. Tokuda, A. W. Black, *Statistical parametric speech synthesis*, Speech Commun., vol. 51, no. 11, pp. 1039-1064, 2009.

[13] T. Drugman, G. Wilfart, T. Dutoit, *A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis*, Proc. Interspeech, Brighton, U.K., 2009.

[14] V. Digalakis, D. Rtischev, L. Neumeyer, *Speaker adaptation using constrained reestimation of Gaussian mixtures*, IEEE Trans. Speech Audio Process., vol. 3, no. 5, pp. 357-366, 1995.

[15] M. Gales, *Maximum likelihood linear transformations for HMM-based speech recognition*, Comput. Speech Lang., vol. 12, no. 2, pp. 75-98, 1998.

[16] J. Ferguson, *Variable Duration Models for Speech*, in Proc. Symp. on the Application of Hidden Markov Models to Text and Speech, pp. 143-179, 1980.

[17] P. Boula de Mareüil, C. d’Alessandro, A. Raake, G. Bailly, M.-N. Garcia, M. Morel, *A joint intelligibility evaluation of French text-to-speech synthesis systems: the EvaSy SUS/ACR campaign*, Proc. LREC, pp. 2034-2037, Gênes, 2006.

[18] C. Benoît, *An intelligibility test using semantically unpredictable sentences: towards the quantification of linguistic complexity*, Speech Commun., vol. 9, no. 4, pp. 293304, 1990.

[19] Noisex-92, Online, *http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html*.

[20] J. Allen, D. Berkley, *Image method for efficiently simulating small-room acoustics*, JASA, vol. 65, no. 4, pp. 943-950, 1979.

The Effects of Frequency-Altered Feedback on the Vocal Productions of Canadian-English Speaking Children

Nichole Scheerer[1], Sarah D'Alton[1], Hanjun Liu[2], Jeffery A. Jones[1]

[1]Department of Psychology and Laurier Centre for Cognitive Neuroscience, Wilfrid Laurier University, Waterloo, Ontario N2L 3C5, Canada
[2]Department of Rehabilitation Medicine, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510080, People's Republic of China
jjones@wlu.ca

Abstract

Speech motor control develops gradually as the acoustics of speech are mapped onto the positions and movements of the articulators. In this study children and adults between 4-30 years of age produced vocalizations while exposed to frequency altered feedback (FAF). The amplitude and latency of the vocal responses were found to differ as a function of age. Furthermore, neurological responses as indexed by the P1-N1-P2 ERP components were also modulated as a function of age. These results suggest that the auditory feedback system undergoes robust changes with age and vocal tract maturation.

Index Terms: speech acquisition, auditory feedback, ERP, P1, N1, P2

1. Introduction

The acquisition of speech production occurs gradually throughout childhood as the acoustics of speech are mapped onto the proper positions and movements of the articulators [1]. This acoustic-motor mapping begins in infancy, when children as young as 6-12 months begin to babble [2]. The auditory feedback produced while babbling allows infants to explore the acoustic consequences of their semi-random articulator movements [3].

Auditory feedback plays an important role in the mapping of speech sounds to articulator positions and movements throughout life. This importance is highlighted by individuals suffering from post-lingual deafness, who display a progressive decline in the vocal quality of their speech production following the loss of auditory feedback [4]. Research has also shown that when people hear their auditory feedback masked [5], or their fundamental frequency (F0) is shifted [6], they make compensatory adjustments to their vocal output. Despite the role of auditory feedback in the monitoring and correction of vocal output, the synaptic and processing delays associated with online monitoring of auditory feedback suggest that other mechanisms also support the production of fluent speech [7].

The state feedback control (SFC) model, a sensorimotor integration circuit based on SFC theory, states that fluent speech is the result of a combination of feedback and feedforward systems [7]. The feedforward system relies on an internal model containing an estimate of the current dynamic state of the vocal tract, based on incoming motor commands as well as previously learned mappings between articulatory movements and the resultant sensory consequences of these movements. Activation of this internal model creates a sensory prediction of the consequences of the upcoming articulatory movements, prior to the arrival of actual sensory feedback. This sensory prediction allows for rapid corrective feedback to be sent to the motor controllers if the actual sensory feedback differs from the expected sensory feedback. In this model fluent speech is primarily driven by the feedforward control system, while actual sensory feedback is reserved for the correction of overt discrepancies between predicted and actual sensory feedback. In addition, actual sensory feedback is used to create and update the mapping between articulatory movements and their resultant sensory consequences [7].

According to the SFC model, fluent speech is highly dependent on the mapping between articulatory movements and the sensory consequences of these movements [7]. Early in speech acquisition these mappings are still being established. In addition, throughout childhood and adolescence these mappings must remain relatively plastic as the shape and size of the vocal tract and articulators are constantly changing. This suggests that as children and adolescents experience developmental changes, they must continually rely on auditory feedback to correct forward sensory predictions.

Despite the increased demands placed on the speech system of children and adolescents as a result of developmental changes, few studies have been conducted to investigate the role of auditory feedback during development. Recently, two behavioural studies were conducted to investigate responses to frequency altered auditory feedback (FAF) in children aged 7-12 years old [1,8]. The results of both of these studies suggest that children have longer latency vocal responses to FAF than adults. Although these results suggest developmental changes in the processing of auditory feedback as a function of age, the narrow range of ages included in these studies makes it difficult to delineate the true influence of age on the processing of auditory feedback.

In the current study, we utilized the FAF paradigm to investigate the behavioural and neurological responses to FAF in children and adults aged 4-30 years. This technique involves randomly shifting the pitch of a speaker's auditory feedback using a digital signal processor [9]. Previous research has shown that when speakers hear their vocal pitch shifted, a reflex-like response occurs that compensates for the pitch alteration by shifting vocal pitch in the opposite direction of the perturbation [10]. We expected that age-related developmental changes would result in different vocal responses to FAF as a function of age. In addition, this study investigated whether development related changes influence the neurological processing of unaltered and FAF. Neurological changes were indexed by the P1-N1-P2 ERP complex, whose components have been shown to differ in amplitude and latency as a function of age.

2. Methods

2.1 Participants

Eighty-two participants were recruited. Participants were divided into 5 separate age groups 4-6 (4 female, 7 male, mean 5.55, SD 0.54), 7-10 (10 female, 10 male, mean 8.51, SD 1.14), 11-13 (7 female, 4 male, mean 11.89, SD 0.83), 14-17 (10 female, 10 male, mean 16.55, SD 0.91), and 18-30 (10 male, 10 female, mean 22.84, SD 2.94) years of age. All participants were native Canadian English speakers, did not speak a tonal language, reported no formal vocal training and were right-handed. All participants received financial compensation or course credit for participation in this study. Informed consent was obtained from each participant, in accordance with the ethical policies at Wilfrid Laurier University.

2.2 Apparatus

Participants were seated in an electrically shielded booth wearing a HydroCel GSN 64 1.0 Cap (Electrical Geodesics, Inc., Eugene, OR) and fitted with Etymotic ER-3 insert headphones (Etymotic Research, Elk Grove Village, IL) as well as a headset microphone (Countryman Isomax E6 Omnidirectional Microphone). During the experiment vocalizations were sent to a mixer (Mackie Oynx 1220, Loud Technologies, Woodinville, WA), followed by a digital signal processor (DSP; VoiceOne, T.C. Hellicon, Victoria, BC), which shifted the pitch of the participant's voice. The pitch-shifted vocalization was then instantaneously presented back to the participant as auditory feedback. The unaltered voice signals were digitally recorded (TASCAM HD-P2, Montebello, CA) at a sampling rate of 44.1 Hz for later analysis.

2.3 Procedure

Participants were seated in front of a computer screen. Vocal utterances were elicited from both the children and adults as part of an interactive space game. The game required that the participants keep their gaze fixed while vocalizing the vowel sound /a/. Participants were instructed to vocalize at a loud, but comfortable amplitude. Vocalizations were played back to the participants in real time via headphones. All participants were instructed to refrain from blinking and making extraneous movements while vocalizing.

The space game elicited 100 vocalizations that were between 4 and 5 seconds in duration. During 50 of these vocalizations, the participant's voice was perturbed -100 cents (down one semi-tone) 3 times per vocalization, with an inter stimulus interval that varied randomly between 1000 and 1200 ms. Each perturbation had a fixed duration of 200 ms. The other 50 vocalizations were unaltered. Although the voice was left unaltered, a sample was taken 3 times per utterance with an inter sample interval that varied randomly between 1000 and 1200 ms for analysis. Altered and unaltered trials were randomly presented throughout the experiment.

2.4 Behavioural Analysis

The digital recording of the vocalizations was segmented into separate utterances. F0 values were calculated for each utterance using the SWIPE' algorithm [11]. F0 values were normalized to a baseline period, 200 ms prior to the onset of the perturbation, by converting Hertz values to cents using the following formula:

$$\text{Cents} = 100(12 \log_2 F/B) \qquad (1)$$

In the formula, F is the F0 value in Hertz and B is the mean frequency of the baseline period.

Cents values were calculated for 200 ms before the perturbation (baseline period), and 500 ms after the perturbation. An average F0 trace was constructed for each shift magnitude, 0 and -100 cents, for each participant.

The amplitude of the participant's compensation response was determined by finding the point at which the participant's average F0 trace deviated maximally from the baseline mean, and the latency was calculated as the time at which this maximal deviation occurred [10].

2.5 EEG Recording and Analysis

EEG data were recorded from 64 scalp electrodes referenced to an electrode at the vertex (Cz). Data were bandpass filtered (1-30 Hz) and digitized (12-bit precision) at 1000 Hz. Electrode impedances below 30 kOhms were obtained prior to the experiment, which ensured impedances were maintained below 50 kOhms throughout the duration of the experiment. After data acquisition, the voltage values were re-referenced to the average voltage across all sites. The data were then epoched into segments from 100 ms before the onset of the perturbation to 500 ms after perturbation onset. Data were analyzed offline for movement artifacts and any segment with voltage values exceeding 55 μV of the moving average over an 80 ms span was rejected.

For each participant, averaged waveforms were created for the 0 cent shifts and the -100 cent shifts for each electrode. Grand average waveforms were created for both conditions by averaging the data from all participants for each electrode, followed by baseline correction. For all average files for each participant, the maximum amplitude and latency was calculated for the ERP components of the P1-N1-P2 complex at the FCz electrode. These components were extracted at time windows from 50-100 ms, 100-200 ms, and 200-300 ms. These windows were chosen based on a visual inspection of the data.

2.6 Statistical Analysis

Repeated-measures analysis of variances (RM-ANOVAs) were conducted comparing the amplitudes and latencies of vocal responses as well as the P1-N1-P2 ERP components across shift, age groups, and sex. In addition, one-way ANOVAs were conducted in order to investigate the influence of age and sex on vocal and ERP response latencies in the shifted condition (-100 cent shift condition).

3. Results

3.1 Behavioural Results

3.1.1 Response Magnitude

A three-way RM-ANOVA was conducted to compare the effects of sex, age, and shift magnitude on vocal response magnitude. The results indicate a significant main effect of shift magnitude ($F(1,72)=84.735$, $p<.001$), as vocal responses were found to be larger to the -100 cent shift relative to the 0 cent shift. A significant main effect of age was also found ($F(4,72)=3.080$, $p=.021$). Bonferroni comparisons indicate that 4-6 year-olds had significantly larger vocal response magnitudes than 7-10 ($p=.015$), 14-17 ($p=.002$), and 18-30 ($p=.010$) year-olds (see Figure 1). The main effect of sex ($F(1,72)=1.202$, $p=.277$), as well as the interaction between sex and age ($F(4,72)=1.010$, $p=.408$), failed to reach significance.

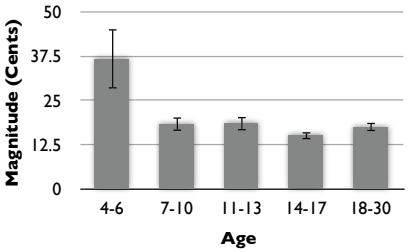


Figure 1: *Behavioural Response Magnitudes.*

3.1.2 Response Latency

A one-way ANOVA was conducted to investigate the effects of sex and age on vocal response latency in the -100 shift condition. The results indicate a main effect of age ($F(4,72)=2.667$, $p=.039$). Bonferroni comparisons indicate that 4-6 year-olds have significantly longer response latencies than 18-30 year-olds ($p=.037$; see Figure 2). In addition, the interaction between age and sex ($F(4,72)=2.517$, $p=.049$) was also significant. Bonferroni comparisons for this interaction however, were not significant. The main effect of sex ($F(1,72)=.131$, $p=.718$) also failed to reach significance.

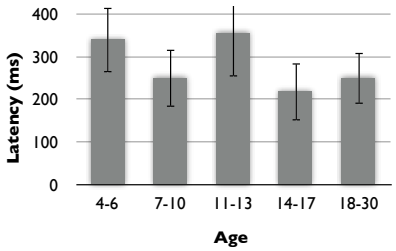


Figure 2: *Behavioural Response Latencies.*

3.2 ERP Results

3.2.1 P1 Amplitude and Latency

A three-way RM-ANOVA was conducted to compare the effects of sex, age, and shift magnitude on the P1 amplitude. The results indicate a significant main effect of shift magnitude ($F(1,72)=64.998$, $p<.001$), as the amplitude of the P1 component was found to be larger to the -100 shift relative to the 0 cent shift. A significant main effect of age was also found ($F(4,72)=6.799$, $p<.001$). Bonferroni comparisons indicate that 4-6 year-olds had significantly larger P1 amplitudes than 14-17 ($p=.004$) and 18-30 ($p=.004$) year-olds, and 7-10 year-olds also had significantly larger P1 amplitudes than 14-17 ($p=.012$) and 18-30 ($p=.009$) year-olds (see Figure 3). In addition, the interaction between shift and age was also significant ($F(4,72)=2.941$, $p=.026$). The main effect of sex ($F(1,72)=.165$, $p=.686$), as well as the interaction between sex and age ($F(4,72)=.897$, $p=.470$), failed to reach significance.

A one-way ANOVA was conducted to investigate the effect of sex and age on P1 latency in the -100 shift condition. The results indicate a main effect of age ($F(4,72)=5.881$, $p<.001$). Bonferroni comparisons indicate that 4-6 and 7-10 year-olds have significantly longer P1 latencies than 18-30 year-olds ($p=.004$ and $p=.001$, respectively). In addition, the main effect of sex ($F(1,72)=7.376$, $p=.008$) was also significant, indicating females ($M=78.85$, $SD=16.52$) have more rapid P1s than males ($M=88.73$, $SD=12.639$). However, the interaction between sex and age ($F(4,72)=1.497$, $p=.212$) failed to reach significance.

3.2.2 N1 Amplitude and Latency

A three-way RM-ANOVA was conducted to compare the effect of sex, age, and shift magnitude on the N1 amplitude. The results indicate a significant main effect of shift magnitude ($F(1,72)=7.434$, $p=.008$), as the amplitude of the N1 component was found to be larger to the -100 shift relative to the 0 cent shift. A significant main effect of age was also found ($F(4,72)=4.203$, $p=.004$). Bonferroni comparisons indicate that 4-6 year-olds had significantly smaller N1 amplitudes than 18-30 ($p=.013$) year-olds (see Figure 3). The main effect of sex ($F(1,72)=.034$, $p=.854$) was found to be non-significant, however the interaction between sex and age ($F(4,72)=2.696$, $p=.037$), was significant. This interaction was due to the fact that 4-6 year-old females ($M=.546$) displayed smaller N1 amplitudes than males ($M=-.609$), while females in all the other age groups 7-10 ($M=-9.60$), 11-13 ($M=-.881$), 14-17 ($M=-1.189$), 18-30 ($M=-1.205$), showed larger N1 amplitudes than males 7-10 ($M=-.227$), 11-13 ($M=-.806$), 14-17 ($M=-.756$), 18-30 ($M=-1.077$).

A one-way ANOVA was conducted to investigate the effect of sex and age on N1 latency in the -100 shift condition. The results indicate a main effect of age ($F(4,72)=4.143$, $p=.004$). Bonferroni comparisons indicate that 11-13 year-olds have significantly longer N1 latencies than 14-17 and 18-30 year-olds ($p=.033$ and $p=.022$, respectively). In addition, the main effect of sex ($F(1,72)=6.364$, $p=.014$) was also significant, indicating females ($M=157.22$, $SD=22.26$) have more rapid N1s than males ($M=168.59$, $SD=16.41$). However, the interaction between sex and age ($F(4,72)=.804$, $p=.527$) failed to reach significance.

3.2.3 P2 Amplitude and Latency

A three-way RM-ANOVA was conducted to compare the effect of sex, age, and shift magnitude on the P2 amplitude. The results indicate a significant main effect of shift magnitude ($F(1,72)=31.546$, $p<.001$), as the amplitude of the P2 component was found to be larger to the -100 shift relative to the 0 cent shift. A significant main effect of age was also found ($F(4,72)=3.451$, $p=.012$; see Figure 3). The main effect of sex ($F(1,72)=.018$, $p=.895$), as well as the interaction between sex and age ($F(4,72)=1.245$, $p=.300$), failed to reach significance.

A one-way ANOVA was conducted to investigate the effect of sex and age on P2 latency in the -100 shift condition. The results indicate the main effect of age ($F(4,72)=1.359$, $p=.257$), sex ($F(1,72)=.068$, $p=.795$), and the interaction between sex and age ($F(4,72)=1.118$, $p=.355$) all failed to reach significance.

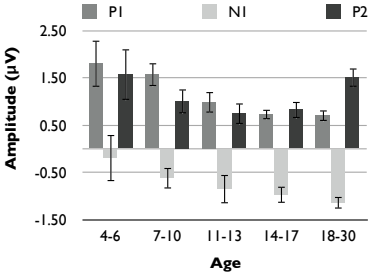


Figure 3: *P1-N1-P2 Response Amplitudes.*

4. Discussion

In this study, children and adults' behavioural and neurological responses to FAF were examined in order to investigate changes in the processing of auditory feedback as a

function of age. The behavioural results indicate that 4-6 year-old children differ from older children and adults in the processing of FAF. Specifically, 4-6 year-old children displayed larger magnitude responses to FAF relative to 7-10, 14-17, and 18-30 year-old participants. Additionally, 4-6 year-old children had slower responses to the FAF relative to 18-30 year-olds, which is in accordance with previous research [1,8]. The larger magnitude responses produced by children in this study may reflect the plasticity of the acoustic-motor mappings in these individuals, as a result of ongoing development. The malleability of children's acoustic-motor mappings may make them more responsive to perceived errors in their speech, causing larger compensations to FAF. On the other hand, adults who possess more robust acoustic-motor mappings, may be more likely to perceive FAF as externally generated, thus causing them to compensate less to the perturbations.

Investigation of the P1 ERP component suggests age and sex related differences in the neurological processing of FAF. More specifically, both 4-6 and 7-10 year-old age groups showed larger amplitude P1 responses than the 14-17 and 18-30 year-old groups, which is consistent with previous findings that the P1 amplitude decreases throughout childhood [12]. In accordance with previous findings, the 4-6 and 7-10 year-olds displayed longer latency P1 responses than the 18-30 year-olds [12,13,14]. In addition to age related effects, P1 responses also differed across the sexes, with males showing longer latency responses than females.

The amplitude and latency of the N1 ERP component also showed differences as a function of age and sex. The N1 peaks elicited by 18-30 year-olds were larger in amplitude than those elicited by 4-6 year-olds. Interestingly, it was the 11-13 year-olds who displayed longer latency responses than the 14-17 and 18-30 year-olds. Similarly to the P1, the N1 responses also differed across the sexes, with males once again displaying longer latency responses than females.

When looking at the P2 ERP component, responses were also affected by age. The amplitude of the P2 component was found to be largest in 4-6 and 18-30 year olds, relative to 7-10, 11-13, and 14-17 year olds. Previous research suggests that the P2 component is modulated by the magnitude of the induced perturbation [15]. Since this experiment only utilized a single shift magnitude, the cause of the observed pattern is currently unclear.

The results of this study indicate that P1 amplitudes decrease with age, while N1 amplitudes increase with age. Previous studies suggest that the P1 ERP component is the childhood correlate of the adult N1 wave. Furthermore, they suggest that the childhood P1 reflects analysis of the basic sound features of a stimulus, while the adults N1 wave reflects higher level sound analysis that is integrated with ongoing mental activity [16]. It is possible that the P1 and N1 results found here reflect a developmental transition from a highly feedback dependent speech system to a feedforward internal model driven speech system.

In addition to amplitude differences, both behavioural and ERP latencies were found to be more rapid in adults than children. It has been suggested that latency changes reflect changes in synaptic density and efficacy in the auditory cortex [17]. Since aging-related changes in the vocal tract and articulators of adults are relatively small compared to developmental changes in childhood, one would expect the synaptic connections in the auditory cortex as well as the pathways between speech related brain regions to remain relatively consistent. As a result of this consistent processing, long-term potentiation is likely to create synaptically efficient connections within the auditory cortex and between speech related brain regions, thus increasing the speed of processing and responding.

5. Conclusions

The age related differences in both behavioural and neurological responses to FAF observed in this study suggest that children and adults process auditory feedback in different manners. This trend demonstrates that the auditory feedback system undergoes robust changes with age and vocal tract maturation. Furthermore, these results suggest that larger response magnitudes and longer response latencies in response to FAF can be used to identify the ongoing development of the audio-vocal system.

6. References

[1] Liu, P., Chen, Z., Larson, C.R., Huang, H., and Liu, H. "Auditory feedback control of voice fundamental frequency in school children," J. Acoust. Soc. Am., 128:1306-1312, 2010.

[2] Pulvermüller, F. and Fadiga, L., "Active perception: sensorimotor circuits as a cortical basis for language", Nature Neurosci., 11:351-360, 2011.

[3] Civier, O., Tasko, S. M. and Guenther, F. H., "Overreliance on auditory feedback may lead to sound/syllable repetitions: simulations of stuttering and fluency-inducing conditions with a neural model of speech production", J. Fluency Disord., 35(3):246-279, 2010.

[4] Waldstein, R. S. "Effects of postlingual deafness on speech production: Implications for the role of auditory feedback," J. Acoust. Soc. Am., 88:2099-2114, 1990.

[5] Bauer, J. J., Mittal, J., Larson, C. R., and Hain, T. C. "Vocal responses to unanticipated perturbations in voice loudness feedback: An automatic mechanism for stabilizing voice amplitude," J. Acoust. Soc. Am., 119:2363-2371, 2006.

[6] Jones, J. A., & Munhall, K. G. "Perceptual calibration of F0 production: Evidence from feedback perturbation," J. Acoust. Soc. Am., 108:1246–1251, 2000.

[7] Hickok, G., Houde, J., and Rong, F., "Sensorimotor integration in speech processing: Computational basis and neural organization," Neuron, 69:407-422, 2011.

[8] Liu, H., Russo, N., and Larson, C.R., "Age-related differences in vocal responses to pitch feedback perturbations: A preliminary study," J. Acoust. Soc. Am., 127:1042-1046, 2010.

[9] Elman, J. L., "Effects of frequency-shifted feedback on the pitch of vocal productions", J. Acoust. Soc. Am., 70:45-50, 1981.

[10] Russo, N., Larson, C., and Kraus, N. "Audio–vocal system regulation in children with autism spectrum disorders," Exp. Brain Res., 188:111-124, 2008.

[11] Camacho, A., and Harris, J.G. "A sawtooth waveform inspired pitch estimator for speech and music," J. Acoust. Soc. Am., 124:1638-1652, 2008.

[12] Oades, R. D., Dittmann-Balcar, A., and Zerbin, D. "Development and topography of auditory event-related potentials (ERPs): mismatch and processing negativity in individuals 8-22 years of age," Psychophysiology, 34(6), 677-693, 1997.

[13] Kraus, N., McGee, T., Carrell, T., Sharma, A., Micco, A., & Nicol, T. "Speech-evoked cortical potentials in children," J. Am. Acad. Audiol., 4(4):238-248, 1993.

[14] Sharma, A., Kraus, N., McGee, T. J., & Nicol, T. G. "Developmental changes in P1 and N1 central auditory responses elicited by consonant-vowel syllables," Electroencephalogr. Clin. Neurophysiol., 104(6):540-545, 1997.

[15] Liu, H., Meshman, M., Behroozmand, R., and Larson., "Differential effects of perturbation direction and magnitude on the neural processing of voice pitch feedback, Clin. Neurophysiol., 122:951-957, 2011.

[16] Ceponiene, R., Rinne, T., and Naatanen, R., "Maturation of cortical sound processing as indexed by event-related potentials," Clin. Neurophysiol., 113:870-882, 2002.

[17] Eggermont, J. J., "The onset and development of auditory function: contributions of evoked potential studies," JSLPA/ROA, 13(1)5:16, 1989.

Can Anybody Read Me?
Motion Capture Recordings for an Adaptable Visual Speech Synthesizer

Simon Alexanderson, Jonas Beskow

KTH Speech Music and Hearing
Royal Institute of Technology, Stockholm, Sweden
{simonal,beskow}@kth.se

1. Introduction

Speech produced in noise exhibits not only increased loudness, but also larger articulatory movements [1]. According to Lindblom's theory of Hyper-Hypo articulation [2], speakers tend to economize their speech production with the goal to make them self understood in a particular communicative situation. For an animated virtual character to function well in different environments, the ability to adapt articulatory effort seems like a useful trait. Below we describe our work towards a visual speech synthesizer capable of simulating articulatory motions for such a character, applicable to different listening conditions. To this end we have used motion capture to record a target speaker trying to make himself understood by a listener, under different conditions: in quiet, in noise and while whispering. The data will later be used to train data-driven articulatory control models for the animated character.

2. Data recording

The speaker was a male Swedish actor who was seated face to face with a listener, and was instructed to read short sentences and words from a monitor, and make sure that the listener understood what was being said. Both listener and speaker wore headphones, where they could hear the their own speech as picked up by a common omni-directional microphone, at a level that was pre-adjusted to roughly to compensate for the attenuation of the headphones. An optional stationary brown noise signal was also fed to the headphones, at different levels throughout the recording (see below).

The speaker's facial movement were recorded by a 10-camera NaturalPoint OptiTrack optical motion capture system operating at 100 frames/sec. The speaker was equipped with 37 reflective facial markers + 4 on the head. In addition, HD-video was captured using a JVC GZ-1 video camcorder. Speech was recorded via a Studio Projects C1 large diaphragm condenser microphone and an RME FireFace 800 external sound card. In order to synchronize the motion capture and the audio, a custom device was constructed, featuring three switchable IR LEDs. When switched on, the LEDs would show up as markers in the motion capture system, at the same time producing an electrical pulse in one input channel of the sound card. The same sync pulses were fed to the external-mic input of the camcorder, thus allow for precise and fully automated post synchronization of all data streams.

The recorded material consisted of 180 short Swedish sentences and 63 nonsense VCV-words (21 Swedish consonants in three different vocalic contexts). A set of 180 English sentences were also recorded.

The full Swedish sentence set was recorded under three different conditions: *Quiet*, *Noisy* and *Whispered*. Quiet is the baseline condition, where no noise was presented in the headphones. In the *Noisy* condition, brown noise at the level of 80 dB SPL was presented in the headphones of both speaker and listener. In the *Whispered* condition, no noise was presented, but the speaker was instructed to keep his voice to a whisper, and still try to make himself understood to the listener. This was done in an attempt to elicit exaggerated lip movements. A reduced set consisting of the 40 first sentences was recorded for two additional noise levels: 70 dB SPL and 90 dB SPL. VCV-words were only recorded in the *Quiet* condition. The English sentence set was only recorded in the conditions *Quiet* and *Noisy* (80 dB).

3. Preliminary data analysis

The motion capture data was sorted and labeled, and cut into segments based on the sync-signal injected by the switched LEDs. A first analysis was made by studying the distance between upper and lower lip. Figure 1 shows this distance for one Swedish sentence, for the quiet and noisy (80 dB) conditions. As expected, the lip movements exhibit much larger amplitude in the noisy conditions than in the quiet.

Table 1 shows that the average inter-lip velocity is lowest in the quiet condition and increases with noise level. The whispered condition exhibits almost twice the velocity as the in the quiet case.

Table 1: Average inter-lip velocity

	Quiet	*70 dB*	*80 dB*	*90 dB*	*Whisper*
Speed mm/s	16.6	26.1	34.0	49.8	32.0

4. Acknoledgement

This work was supported by EU Lifelong Learning Programme, project LipRead (http://www.lipread.eu)

5. References

[1] Fitzpatrick, M., Kim, J. & Davis, C. (2011): "The effect of seeing the interlocutor on auditory and visual speech production in noise", In *AVSP-2011*, 31-35.

[2] Lindblom, B. (1990): *Explaining phonetic variation: A sketch of the H & H theory,* In W. J. Hardcastle & A. Marchal: "Speech Production and, 403-439, Kluwer Academic Publishers, Dordrecht.

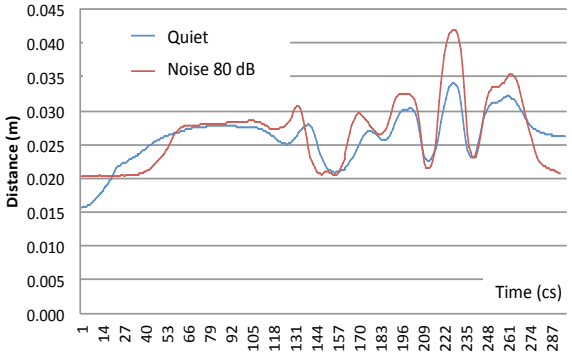


Figure 1: Intrer-lip distance for the Swedish sentence *Dom flyttade möblerna.*

MAGE : A Platform for Performative Speech Synthesis New Approach in Exploring Applications Beyond Text-To-Speech

Maria Astrinaki, Nicolas d'Alessandro, Thierry Dutoit

numediart - Institute of New Media Art Technology, University of Mons, Belgium
maria.astrinaki@umons.ac.be, nicolas@dalessandro.be, thierry.dutoit@umons.ac.be

Speech is the richest and most ubiquitous modality of communication used by human beings. Voice production is one of the most expressive instruments of the human body and people have always been fascinated for artificial speech and singing production. Even though vocal behavior and expression is a very complex mechanism, we realize a highly interactive and social process. Until now, artificial voice production has been based on the Text-To-Speech (TTS) technologies, converting a static text into an intelligible and natural waveform, lately with great success. However the relevance of speech production is not only based on these properties, it is also linked to the context, to an ongoing process of interaction between speakers and to their social environment and background. Through this interdisciplinary redefinition of expressive speech, we see that the nature of voice production is primarily a realtime, dynamic gesture, involving the vocal organs, face and body. Speech is a performance, a gestural phenomenon that transmits messages with both information and emotions.

1. The MAGE Synthesis Platform

As these new trends in understanding expressivity in speech are being explored, one might notice that a real solid platform for performative speech synthesis is missing. Indeed TTS, as a platform, has been tackling and greatly solving the problem of text reading, but not the problem of the artificial speaker. In the text reading paradigm, a significant amount of text is required in advance to be processed into sound as a whole target. Moreover, during this text to speech conversion process any external influence is rather limited, since it results in sound quality degradation. On the other hand, an artificial speaker enables interactive communication by inferring on speech outputs at various production levels and time scales. In fact, such a system has totally different requirements. It needs to have a reactive programming architecture, and to be both listener-specific and context-aware. To our current knowledge, MAGE is the first platform for reactive programming of speech synthesis solutions. It was released recently as a C/C++ open-source project [1, 2].

The MAGE[1] speech synthesis engine is based on HTS [3], the open-source HMM-based statistical parametric synthesis algorithm from Tokuda et al [4]. While still not reaching the intelligibility and naturalness of non-uniform unit selection (NUU) algorithms, HMM-based approaches are quickly improving and, more importantly, they are based on a highly flexible well-defined architecture. Indeed both HMM-based trajectory generation and MFCC-based sound synthesis can be deeply modified. MAGE is a complete architectural redesign of HTS, streaming the speech sound in realtime, according to synthesis parameters that are sent on-the-fly. MAGE has a high-quality output while rendering incoming labels, durations, pitch and vocal tract length parameters with only one phonetic-label delay.

[1]The MAGE project is funded through two PhD grants by the University of Mons (numediart, grant 716631) and Acapela Group S.A.

2. Implementation and Integration

MAGE provides an API for reactive programming in C and C++, aimed at being included in realtime audio softwares. It is thread safe and independent from the actual HTS-based speech engine. In this version, we release MAGE with the pHTS engine, a performative modification of HTS. Fast and easy prototyping is possible with MAGE, since it can be easily imported into other platforms. It can be simply combined with OSC-enabled sensors, allowing both context and prosody user control. Contextual control is implemented by parsing on-the-fly small groups of phonemes, that we can "chunks" into streams of phonetic labels. Prosody control is based on altering on-the-fly pitch curves, duration or vocal tract parameters from gestures.

MAGE comes as a consequential implementation, following the idea of performative speech synthesis, as a way of looking beyond TTS. It results from discussions across different disciplines, such as speech processing, linguistics and human-computer interaction (HCI), attempting to bring a new platform for addressing their problems. In the area of computer music and new interfaces for musical expression, MAGE is targeted to combine simple prototyping with meaningful gestural control, to bring synthetic speech on stage and suggest new art forms.

3. Discussion

Our goal is to build a new framework for understanding long-term questions in speech production, such as degrees of coarticulation, speech motor control, speech planning, intonation, voice quality, speech time scales, etc. through gestural control and interactive interfaces, mainly through mobile and social computing. At this point it is important that we have feedback for this project from actual users. Users that set to work this library, have questions, suggestions and comments. Convening the expectations of the users towards more interactive use of speech synthesis, will eventually bring the next release of the platform to address this practical feedback and become a better tool for engineers, linguists and artists.

4. References

[1] http://sourceforge.net/p/magephts

[2] http://www.numediart.org/demos/mage_phts

[3] K. Tokuda, H. Zen, J. Yamagishi, A. Black, T. Masuko, S. Sako, T. Toda, T. Nose and K. Oura, The HMM-based speech synthesis system (HTS), 2008, http://hts.sp.nitech.ac.jp/

[4] H. Zen, K. Tokuda and A. W. Black, "Statistical Parametric Speech Synthesis," Speech Comm., vol. 51, pp. 1039–1064, 2009.

Overlap behaviour in task-oriented dialogue

Vincent Aubanel[1,2], *Martin Cooke*[1,2], *Catherine Mayo*[3] *and Robert Clark*[3]

[1]Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
[2]Ikerbasque (Basque Foundation for Science)
[3]Centre for Speech Technology Research, University of Edinburgh
v.aubanel@laslab.org

Abstract

Speakers change the way they speak depending on the surrounding environment. When masking noise obstructs the communication channel between interlocutors, they consistently engage in Lombard speech, whose spectral characteristics are well described (e.g. [1, 2, 3]) and are believed to result in better intelligibility through energetic masking reduction (e.g. [4]). However, less is known about how speakers adapt to the temporal characteristics of a fluctuating masker, and whether any such changes aid communication.

In this study pairs of speakers were recorded while engaged in a sudoku-solving task in quiet and in several masking conditions. Maskers were either a competing talker or speech modulated noise with identical temporal characteristics, chosen to investigate the informational masking potential of the masker. The silence density of each masker was also manipulated by adjusting the durations of pauses in the masker to 33% or 66% of the overall duration of the masker.

In all masking conditions, speakers displayed a reduction of overlap with the masker relative to a baseline computed from the masker and speech produced in quiet (Figure 1). The overlap reduction tended to be larger for less dense maskers, and was obtained, at least for the competing talker case, in spite of an increase of speech activity in the masking condition. These results confirm [5], and indicate that speakers were able to exploit the temporal silences of the masker to convey information necessary to complete the task.

While denser maskers led to stronger Lombard effects, such as greater increase of $F0$ and $F1$ compatible with the energetic masking potential of the masker, less dense maskers induced a qualitatively different behaviour in speakers, which we characterise as a "wait and talk strategy", demonstrated by a decrease of speaker onsets following masker onsets and an increase of these events following masker offsets.

Taken together, these results suggest that talkers respond to their environment by making speech modifications which have the potential to help communicate with their interlocutors.

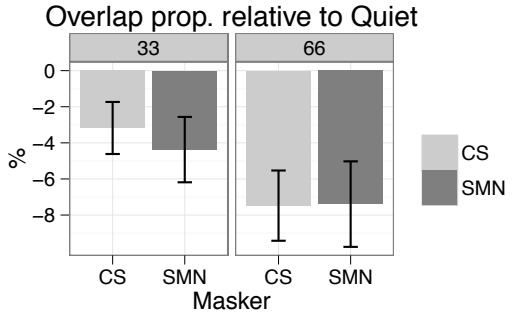


Figure 1: Overlap reduction relative to quiet in two silence densities: (33% and 66%), for a masker of competing speech (CS) and speech modulated noise (SMN).

Acknowledgements. This work was supported by EU Future and Emerging Technology (FET–OPEN) Project LISTA (The Listening Talker).

References

[1] W. Van Summers, D. B. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[2] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, 1993.

[3] M. Garnier, "Communiquer en environnement bruyant : de l'adaptation jusqu'au forçage vocal ," Ph.D. dissertation, Université Paris 6, 2007.

[4] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3261–3275, 2008.

[5] M. Cooke and Y. Lu, "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2059–2069, 2010.

Lombard and temporal effects in concurrent conversations

Vincent Aubanel[1,2], Martin Cooke[1,2], M Luisa Garcia Lecumberri[1], Catherine Mayo[3] and Robert Clark[3]

[1]Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
[2]Ikerbasque (Basque Foundation for Science)
[3]Centre for Speech Technology Research, University of Edinburgh
v.aubanel@laslab.org

Abstract

Conversing in the presence of a background conversations is an everyday act yet little is known about how speakers maintain intelligibility and comprehensibility when confronted with a background of intelligible speech. Speaking in non-informative noise (Lombard speech) has largely been described in terms of induced spectral changes, but it is possible that speakers employ a richer set of strategies, including temporal modifications, to help overcome the disrupting effect of competing speech (e.g. [1]).

In the current study pairs of British English talkers engaged in natural dialogues in the presence or absence of another talker pair. Talkers were instructed to converse only with the other interlocutor in their pair. Pairs sat facing each other around a round table, so that when both pairs were present, talkers had to "talk across" the other pair. In half of the conditions talkers wore visors which prevented them seeing their interlocutors but with no effect on audio transmission.

In both face-to-face and audio-only conditions, speaking simultaneously with another talker resulted in overall increases in energy, F0, F1 and a decrease in speech rate. A distinction between within- and across-pair overlaps however revealed that overlapping with the background pair resulted in an increase in energy but no change in the two prosodic parameters F0 and speech rate, whereas within-pair overlap led to an increase in F0 and a decrease in rate, and no change in speech level (Figure 1). This contrasts with previous studies on simultaneous conversations where the seating configuration did not demand "talking across" the other pair [2], suggesting that background "noise" that consists of intelligible speech does not automatically induce increases in speech output level routinely observed when speaking in the presence of, for example, stationary noise.

Not seeing the interlocutor actually led to a decrease of energy during within-pair overlaps, contrasting with [3] where more effortful speech was oberved, however with non-interactive maskers. This absence of visual cues also led speakers to reduce their overlap with their interlocutor, and to a greater extent when the background

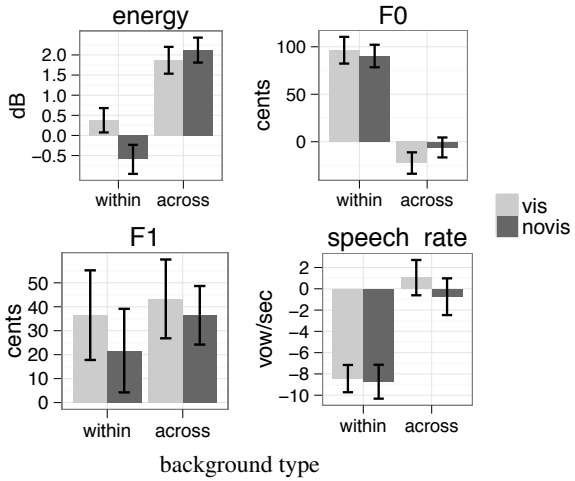


Figure 1: *Lombard effects contrasting within- and across-pair overlaps background types.*

pair was present. Although overlap with a background speaker was as high as 80%, we also uncovered evidence of turn-taking behaviour between foreground and background speakers, hinting at a speaker overlap avoidance strategy, albeit necessarily rather weak in such a dense speech background.

Taken together, the finding suggests that adverse conditions cause interlocutors to adopt more careful dialogue strategies, perhaps to reduce both energetic and informational masking at the ears of the listener.

Acknowledgements. This work was supported by EU Future and Emerging Technology (FET–OPEN) Project LISTA (The Listening Talker).

References

[1] M. Cooke and Y. Lu, "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2059–2069, 2010.

[2] V. Aubanel, M. Cooke, J. Villegas, and M. L. Garcia Lecumberri, "Conversing in the presence of a competing conversation: effects on speech production," in *Interspeech*, Florence, Italy, 2011, pp. 2833–2836.

[3] M. Fitzpatrick, J. Kim, and C. Davis, "The effect of seeing the interlocutor on speech production in different noise types," in *Interspeech*, Florence, Italy, 2011, pp. 2829–2832.

How Should Attentive Speaker Agents Adapt to Listener Feedback?

Hendrik Buschmeier, Stefan Kopp

Sociable Agents Group, CITEC and Faculty of Technology, Bielefeld University
PO-Box 100131, 33501 Bielefeld, Germany
{hbuschmeier,skopp}@uni-bielefeld.de

Introduction In dialogue, a speaker's communicative actions are not only shaped by her communicative intentions and her ability to express them linguistically, but importantly also by the actions and reactions of her interlocutor, which she – being cooperative – cannot ignore. Both dialogue partners contribute to dialogue success and collaborate on the interaction to make it as efficient as possible by responding to each other's needs.

One important mechanism for dialogue coordination is *communicative feedback* in the form of short verbal-vocal expressions (such as 'uh-huh', 'yeah', 'huh?'), head movements (nods, shakes, etc.), facial expressions (e.g., smiling, frowning, raising an eyebrow) and gaze. The use of feedback is prevalent in spoken interaction. Listeners often produce it concurrently to the speaker's communicative actions and convey that they are in contact with the speaker, whether they perceive and understand what the speaker says, whether they accept, adopt or agree with the speaker's utterance and also further attitudes towards it [1].

By providing feedback, a listener thus reveals parts of his mental state and indicates in a timely manner how the interaction is going or which attitude he has towards an utterance. On the basis of feedback, a speaker can then reason about the listener's mental state and use this information to adapt her subsequent utterances to the listener's needs. If, for example, the listener frowns right after the speaker mentions an object which the speaker thinks she unambiguously referred to, she can use this evidence of difficulties of understanding and clarify the reference by providing additional information.

Attentive speaker agents Currently, artificial conversational agents (such as dialogue systems or embodied virtual agents) lack capabilities to deal with user feedback. This is one aspect why interacting with them is often cumbersome. Users are forced to communicate meta-information on the state of the conversation explicitly and adhere to strict turn-taking behaviour while doing it. And if agents can react to this information at all, they do so no sooner than in their next utterance.

In previous work [2], we proposed that conversational agents should be *attentive speakers*, which we define as being able to (i) elicit feedback from users; (ii) detect and interpret concurrent user feedback; and (iii) respond to feedback by adapting their conversational actions to accommodate the user's needs. Such an attentive speaker agent can determine problems as soon as they become evident and is thus able to respond immediately by adapting the still unspoken part of its current utterance.

In the above mentioned work, we also presented a first approach towards conversational agents that can attend to and adapt to communicative listener feedback. The agent, which assists users in organising their weekly calendar, attributes a simple numerical model of *listener state* ($C, P, U, A, dU, dP \in [0,1]$) to the user. These values are updated when feedback signals (head gestures, simple feedback expressions, gaze) are encountered. Based on this *attributed listener state*, the agent's incremental natural language generation component then changes parameters and constraints that shape the form of the unspoken increments of the utterance. When users show difficulties understanding what the agent means, for example, redundancy is introduced by making implicit communicative effects explicit.

Corpus analysis Here, we present first results from a dialogue study of human–human interactions in the calendar domain. We analyse the semantic and pragmatic properties of listeners' feedback signals as well as speakers' utterances in their vicinity.

Listeners' feedback signals are annotated on multiple dimensions. We classify them according to their basic communicative function and also look for signs of uncertainty, progressiveness and attitude that are often conveyed. We use this information to reason about a listener's mental state with respect to the utterance a feedback signal refers to.

Speakers' utterances, on the other hand, are analysed for their illocutionary force as well as for grounding status and information state of their content (e.g., is it new and possibly unexpected or already known to both interlocutors). The utterance parts succeeding listers' feedback signals are further analysed with respect to the parts preceding them. We do this in order to find out whether new information is introduced and what role it plays; whether old information is clarified or used redundantly; whether implicit content is made explicit; etc.

The insights gained from this corpus analysis will be used to inform the design of a Bayesian model of the listener that takes the speaker's utterances, contextual factors as well as the listener's feedback signals into account when reasoning about the attributed listeners state [3]. The analysis of feedback-succeeding utterance parts will help us identify and implement additional adaptation mechanisms and strategies for the natural language generation component of the attentive speaker agent.

Acknowledgments This research is supported by the Deutsche Forschungsgemeinschaft (DFG) at the Center of Excellence in 'Cognitive Interaction Technology' (CITEC).

References

[1] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, pp. 1–26, 1992.

[2] H. Buschmeier and S. Kopp, "Towards conversational agents that attend to and adapt to communicative user feedback," in *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, Reykjavik, Iceland, 2011, pp. 169–182.

[3] ——, "Unveiling the Information State with a Bayesian model of the listener," in *SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue*, Los Angeles, CA, 2011, pp. 178–179.

Controlling Voice Source Parameters to Transform Characteristics of Synthetic Voices

João P. Cabral, Julie Carson-Berndsen

School of Computer Science and Informatics, University College Dublin, Ireland

joao.cabral@ucd.ie, julie.berndsen@ucd.ie

1. Introduction

The success of the speech communication process depends not only on the intelligibility of the speech transmitted to the listener but also on how the message is spoken. An important aspect that carries underlying information in speech besides the linguistic content is the type of voice. Humans change instinctively or intentionally their voice depending on their mood, the environment, the listener, the feelings they want to transmit, etc. In order to take advantage of this valuable aspect of speech in applications of synthetic voices for spoken communication is necessary that the computer can produce a high variability of voices and that it can predict an appropriate voice based on environmental cues, including feedback information about the listener. This work fits into the subject of modelling and transforming acoustic aspects of speech for controlling the type of synthetic voice. The goal is to accurately model an important acoustic component of speech related to voice characteristics which is aspiration noise. This noise signal results from the turbulence of air passing through the glottis during human speech production. It can be represented by an amplitude modulated Gaussian noise, which depends on the glottal volume velocity and glottal area. For example, this modulation effect is more important in breathy voice than modal because the vocal folds usually do not completely close for breathy unlike modal.

Index Terms: Voice transformation, aspiration noise, breathy

2. Voice Transformation Method

2.1. Glottal Spectral Separation for Analsyis-Synthesis

In this work, the analysis-synthesis method called Glottal Sepectral Separation (GSS) [1] is used to transform a modal voice into breathy. This technique was chosen because it permits to control parameters of an acoustic glottal source model, the Liljencrants-Fant (LF) model [2], and performed well in voice transformations. However, the LF-model does not represent the noise characteristics of the voice source, in particular the aspiration noise. For this reason, we combined the GSS method with a technique for modelling this type of noise.

2.2. Aspiration Noise Modelling

The aspiration noise was estimated using an harmonic-stochastic model of speech (HNM) The harmonic and stochastic components were firstly separated from the speech signal using the UPC tools (`http://www.talp.upc.edu/talp/index.php/resources/tools/`). Then, the harmonic signal was subtracted from the speech signal to obtain the noise component. Finally, the aspiration noise was estimated from the noise by LPC inverse filtering.

After estimating the aspiration noise it is necessary to model its amplitude modulation effect. The modulation function is calculated by using the Hilbert transform method of envelope detection. In this work, this envelope is initially parameterised using a polynomial fitting technique. Then, a triangular function is obtained from the polynomial representation in order to obtain a more robust and accurate estimate of the noise envelope. A great advantage of the triangular representation compared to typical functions that have been used to represent the envelope, such as the symmetric Gaussian and Hanning windows, is that it better represents different shapes (including asymmetric shapes) of the energy envelope. This flexibility is similar to that of using a glottal source signal representation of the energy envelope [3], but the first is simpler because does not require estimation of the glottal signal. For example, by using this function is possible to adjust the envelope shape depending on the transformations of the glottal parameters.

3. Experimental Results

The values of the voice quality parameters of the LF-model (open quotient, speed quotient and return quotient) calculated for an utterance (modal voice) were modified to obtain parameter contours with mean values equal to those measured on the target voice, similarly to [1]. We additionally modified the mean values of the F_0 and HNR contours to improve the voice conversion. HNR was estimated as the energy ratio between the harmonic and noise components of the HNM.

Results of a perceptual experiment showed that the GSS method combined with the aspiration noise model significantly improved the naturalness of synthetic speech and transformation of modal voice into breathy, compared with the baseline GSS method which only used the LF-model to represent the excitation signal.

4. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`www.cngl.ie`) at University College Dublin.

5. References

[1] Cabral, J. P., Renals, S., Richmond, K. and Yamagishi, J., "Glottal Spectral Separation for Parametric Speech Synthesis", Proc. of INTERSPEECH, 1829–1832, Brisbane, 2008.

[2] Fant, G., Liljencrants, J. and Lin, Q., "A four-parameter model of glottal flow", STL-QPSR, 26(4), 1–13, 1985.

[3] Degottex, G., Rbel, A., Rodet, X., "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter", Proc. of ICASSP, 5128–5131, 2011.

Listening talkers produce great spectral tilt contrasts

Thomas Ulrich Christiansen[1], *Jan Heegård*[2], *Peter Juel Henrichsen*[2]

[1]Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark, Lyngby, Denmark
[2]Department of International Language Studies and Computational Linguistics, Copenhagen Business School, Copenhagen, Denmark

tuc@elektro.dtu.dk, jhp.isv@cbs.dk, pjh.isv@cbs.dk

Abstract

It is well known that the envelope of the long-term average speech spectrum flattens with vocal effort. A recent study [1] showed that content words had a flatter spectral envelope than content words at the same overall level for a specific Danish speech material.

The present paper investigates whether this effect is present in a larger and more diverse speech material, and if the effect is greater when the talker is listening (participating in a dialogue) as compared to monologue.

The monologue speech material consisted of recordings from 18 native talkers of Danish describing a network of colored geometrical shapes taken from DanPASS [2]. The spectral tilt was gauged by calculating the band-level difference in dB between two frequency bands with pass-bands 150 to 803 Hz and 803 to 1358 Hz respectively in 5 ms intervals.

This was done separately for intervals containing content words and function words and grouped by talker. The spectral tilt difference was then calculated as the average band-level difference for function words minus the average band-level difference for content words. This calculation was grouped per talker. For the monologues these differences ranged between 5 and 8 dB for the 18 talkers.

Content words were defined as nouns, active verbs, adjectives and adverbs. Function words were defined as articles, pronouns, conjunctions and auxiliary verbs. Words not belonging to any of these categories were not used.

The dialogue speech material was also from DanPASS and consisted of recordings from 13 of the same talkers as the monologues. In the dialogue speech material talkers where asked to describe a map with certain discrepancies and negotiate their way through the map.

Spectral tilt differences between content- and functions words were calculated in the same way as for the monologues. The results show that the spectral tilt differences are slightly higher for dialogues than monologues. A two-way anova (grouped by talker and word type) showed that these differences are significant.

We conclude that Danish talkers mark high information density in spontaneous speech (=content words) by means of flat spectral envelope, not just for monologues, but also for dialogues. Moreover, when engaged in dialogue, talkers enhance this spectral flattening.

In our view it is remarkable that conclusions with statistical validity can be reached based on the over-simplified definition of spectral tilt employed in this paper. We speculate that optimizing both the definition of spectral tilt and the word categories comprising content- and function words, may allow us to observe even greater effects than reported here.

The eventual goal of this line of research is to devise a simple, tractable method for distinguishing high information content from low information content in speech, based on the ubiquitous assumption that content words carry more information than function words. Such a method could potentially be applied in hearing aids, cochlear implants and automatic speech recognition.

Index Terms: spectral tilt, spectral envelope, speech production, speech perception, content words, function words.

References

[1] Henrichsen, P.J. and Christiansen, T.U., ”Information based speech transduction”, Proceedings of 3rd International Symposium on Auditory and Audiological Research, August 2011, Nyborg, Denmark, in press, 2011

[2] Grønnum, N., “A Danish phonetically annotated spontaneous speech corpus (DanPASS)”, Speech Communication 51, 594-603,2009

Do non-native listeners benefit from speech modifications designed to promote intelligibility for native listeners?

Martin Cooke[1,2], *Maria Luisa Garcia Lecumberri*[2], *Yan Tang*[2], *Mirjam Wester*[3]

[1]Ikerbasque (Basque Science Foundation), Bilbao, Spain
[2]Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
[3]Centre for Speech Technology Research, University of Edinburgh, Edinburgh UK

1. Motivation

Given the increasing use of sound output technologies in cluttered acoustic environments such as transport interchanges, it is of interest to discover ways to modify generated speech which maximise the likelihood of correct message reception, without resorting to excessive sound output level or repetitive announcements. Additionally, speech modifications deployed in public spaces should ideally be beneficial to all listeners, regardless of their first language.

2. Design

We compared the effect of a number of speech modification strategies in noise on native and non-native listeners. All were based on energy reallocation across time and frequency under constant SNR and duration constraints [1, 2]. British English and Spanish listeners identified letter and digit keywords in short sentences for natural speech (**original**) and for six types of modification:

SegSNR frame-wise SNR equalisation

ChanSNR frequency channel SNRs equalised

LocalSNR SNRs equalised in time-frequency region

SelectBoost energy reallocation to boost just-audible time-frequency regions

Pausing introduction of a short pause to avoid most intense noise epoch

Combined **SelectBoost** + **Pausing**

Speech was presented mixed with either stationary speech-shaped noise (SSN) or speech envelope modulated noise (SMN) at a global SNR of -6 dB.

3. Results

While native and non-native listeners' absolute scores differed by about 14 percentage points, the pattern of scores as a function of modification type was strikingly similar, with a correlation of 0.97 (see figure 1). For example, both listener groups benefitted most from selective boosting of regions close to threshold, and both were adversely affected by pausing in the presence of stationary noise.

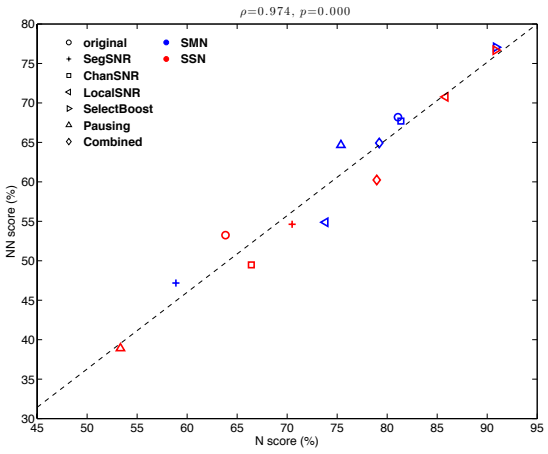


Figure 1: *Native and non-native scores for natural and modified speech in stationary and fluctuating noise.*

4. Discussion

Speech signal modifications exist which are equally effective for native and non-native listeners, suggesting that these modifications result in language-independent auditory changes which may be effectively transferred to other target languages and listener populations.

Acknowledgement. This study was supported by the LISTA Project FET-Open grant no. 256230.

5. References

[1] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 1636–1639.

[2] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 345–348.

Identifying Tenseness of Lombard Speech Using Phase Distortion

Gilles Degottex[1], *Elizabeth Godoy*[2] *and Yannis Stylianou*

University of Crete, Computer Science Dep. and FORTH, Inst. of Computer Science
Vasilika Vouton, 71110 Heraklion, Greece

Abstract

The "Lombard effect" describes speakers' tendency to increase their vocal effort when communicating in noise [1]. Most often, the Lombard effect is examined in terms of acoustic parameters such as pitch, duration, and spectral amplitude (e.g., tilt/slope, formants) [2, 3]. However, these parameters offer limited insight into voice quality, such as "tenseness" associated with increased vocal effort. Acoustically, one of the most significant indicators of tenseness relates to features of the glottal excitation signal: specifically, perceived tenseness of a voice is linked to a decrease in the glottal spectral tilt (i.e., slope) [4]. Unlike typical analyses of the Lombard effect, the work in [5] explicitly examines glottal source parameters. Unfortunately, glottal source estimation is a challenging and delicate problem. Consequently, the present work seeks to offer an alternative analysis framework that can also isolate contributions of the excitation source (from the vocal tract), but without explicit glottal source modelling. In particular, phase distortion (defined below) is used to highlight differences in voice quality, focusing specifically on the relative tenseness of Lombard speech compared to Normal speech.

The phase distortion is defined here as the group-delay with the influence of linear phase component removed. To calculate it, a harmonic model with pitch-synchronous analysis [6] is used. First, the minimum phase contribution of the vocal tract filter is removed from the instantaneous phase of the sinusoidal parameters. Then, using phase difference and antidifference operators, the linear phase component is also removed from the measurement. What remains after this processing is phase distortion, which depends only on the shape of the glottal signal. This computation has been already proposed for the estimation of glottal model parameters and emotional valence detection [7, 8]; that is, the link between this phase distortion and voice quality has been established. In the context of Lombard speech analysis, the variance of this phase distortion (calculated across 2 pitch periods and averaged with an ERB scale) is more interesting than the phase distortion itself, since it reveals the stability of the sinusoidal components in the time-frequency plan. Specifically, the boost of glottal source energy above aspiration noise levels at high frequencies (in voiced segments) can be clearly observed by examining the phase distortion variance (see Figure 1).

Figure 1 shows an example of aligned Normal and Lombard speech, along with the corresponding spectrograms and phase distortion variance. First, note that for the unvoiced,"noisy" parts of both the Normal and Lombard speech, the phase distortion variance is large for all frequencies. Now, considering the voiced speech segments, it can be seen that the phase distortion variance is large for the upper part of the spectrum (high frequencies) of the Normal speech and small for lower frequencies. On the other hand, for the Lombard speech, the phase distortion variance remains relatively low across all frequencies for the voiced segments. This effect can be seen most clearly comparing the segments before 0.4 and after 0.8 seconds. Referring back to previous discussions, the lower phase distortion variance indicates the boosting of glottal spectral energy above aspiration noise for tense voices. Thus, the phase distortion variance is effectively confirming the tenseness in the Lombard speech.

In addition to identifying tenseness in Lombard speech, the framework for the present analyses could potentially be extended to synthesize speech with different voice quality. That is, we observe the tenseness of Lombard speech in relation to decreased phase distortion variance. So, by reducing the phase distortion variance, can we synthesize a voice that is more tense (admittedly, a difficult task)? Furthermore, in terms of the Lombard effect, the present work clearly shows differences in voice quality between the Normal and Lombard speech, but these observations beg the question: how do differences in voice quality influence speech intelligibility?

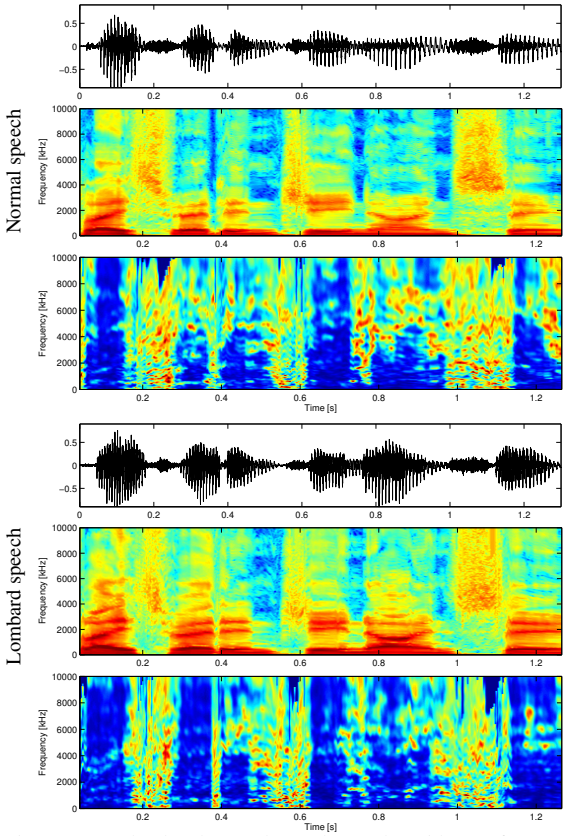


Figure 1: Lombard and normal speech samples with waveform, spectrogram and phase distortion variance

[1] Thanks to Swiss National Science Foundation (PBSKP2_134325).
[2] Thanks to LISTA project (E.U. FP7 FET-OPEN, 25623).

1. References

[1] Etienne Lombard, "The sign of the elevation of the voice," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 2, pp. 101–119, 1911.

[2] J.C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *JASA*, vol. 93, no. 1, pp. 510–524, 1993.

[3] W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech producton: Acoustical and perceptual analyses," *J. Acous. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[4] G. Fant, "The LF-model revisited. transformations and frequency domain analysis.," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[5] T. Drugman and Thierry Dutoit, "Glottal-based analysis of the lombard effect," in *Interspeech*, 2010, pp. 2610–2613.

[6] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 290–300, 2010.

[7] G. Degottex, A. Roebel, and X. Rodet, "Function of phase-distortion for glottal model estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 4608–4611.

[8] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features and glottal features for emotional valence detection," in *Proc. International Conference on Speech Prosody*, 2012.

Simple Spectral Techniques to Enhance the Intelligibility of Speech using a Harmonic Model

Daniel Erro[1], *Yannis Stylianou*[2,1], *Eva Navas*[1] *and Inma Hernaez*[1]

[1]Aholab Signal Processing Laboratory, University of the Basque Country (UPV/EHU), Bilbao, Spain
[2] Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
{derro,eva,inma}@aholab.ehu.es, yannis@csd.uoc.gr

Abstract

We have designed a tool to increase the intelligibility of speech by manipulating the parameters of a harmonic speech model. The system performs the transformation in two steps. In the first step, it modifies the spectral slope, which is closely related to the vocal effort. In the second step, it amplifies low-energy parts of the signal using dynamic range compression techniques. Such a system has two main advantages: its simplicity and the fact that it can be easily integrated into the synthesis engine of a speech synthesizer trained from Mel-cepstral coefficients.

1. Introduction

Speech synthesizers are usually trained from clean speech databases recorded by professional speakers in silent environments. Consequently, when synthetic speech is played in noisy conditions it is often hard for listeners to understand the message. For speech synthesizers to be practical in different contexts, it is desirable to have some control over the voice characteristics that play a crucial role in intelligibility.

There are basically two ways of modifying the synthetic speech to make it more intelligible in noisy conditions: (i) acquiring the right databases and then using statistical mapping techniques [1]-[4]; (ii) using expert knowledge and signal processing techniques to enhance the output of the synthesizer [3][5]. This work follows the latter approach, which is advantageous in the sense that it avoids recording new databases and retraining the underlying models.

In this paper we show that the parameters of the harmonic model can be modified to increase speech intelligibility in noise. Since the harmonic model is applicable to synthesizing high-quality speech from a statistically generated Mel-cepstral vector sequence [6], the proposed transformation can be easily integrated into the synthesizer, which makes it able to modify its output according to the environmental noise at the expense of an almost negligible increment of the computational load.

2. Brief Description

The harmonic model assumes that locally stationary speech signal segments can be decomposed into a series of harmonically related sinusoids represented by their frequencies, amplitudes and phases. The proposed system operates entirely on the parameters of this model. It consists of two transformation steps to be applied in cascade.

Lombard speech is characterized by a higher vocal effort, which has an impact on the spectral tilt. We have studied a simple spectral transformation that modifies the amplitudes by a constant slope measured in dB/decade and then renormalizes their energy at frame-level.

Under the hypothesis that nonstationary portions of the speech signal (plosives, for instance) play a decisive role in intelligibility, the second transformation studied in this work aims at amplifying low-energy frames, which are likely to contain most of these meaningful portions. We have explored modification procedures inspired by Dynamic Range Compression (DRC) techniques, which can also be implemented in the harmonic amplitude domain.

3. Results

Objective tests based on our implementation of the extended Speech Intelligibility Index (SII) [7] indicate that the intelligibility of natual signals in speech-shaped noise improves significantly after manipulation. The two steps of the transformation contribute cumulatively to this improvement, each one increasing the SII score by approximately 0.1. Formal subjective evaluations carried out by 78 native listeners confirm that the enhancement system reduces the median word error rate at sentence level by 30-50% for $-9\text{dB} \leq \text{SNR} \leq 1\text{dB}$.

4. Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and the Basque Government (ZURE_TTS Project).

5. References

[1] B. Langner, A. Black, “Improving the understandability of speech synthesis by modeling speech in noise”, Proc. ICASSP, 2005.
[2] B. Picart, T. Drugman, T. Dutoit, “Continuous control of the degree of articulation in HMM-based speech synthesis”, Proc. Interspeech, 2011.
[3] T. Raitio, A. Suni, M. Vainio, P. Alku, “Analysis of HMM-Based Lombard Speech Synthesis”, Proc. Interspeech, 2011.
[4] Z.H. Ling, K. Richmond, J. Yamagishi, R.H. Wang, “Integrating articulatory features into HMM-based parametric speech synthesis”, IEEE Trans. Audio, Speech & Lang. Process., 2009.
[5] D.Y. Huang, S. Rahardja, E.P. Ong, “Lombard Effect Mimicking”, Proc. 7th ISCA Speech Synthesis Workshop, 2010.
[6] D. Erro, I. Sainz, E. Navas, I. Hernaez, “Improved HNM-based Vocoder for Statistical Synthesizers”, Proc. Interspeech, 2011.
[7] K.S. Rhebergen, N.J. Versfeld, “A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners”, J. Acoust. Soc. Am., 2005.

Glissando Dialogs: a Corpus for the Analysis of Entrainment in Phone Services

David Escudero[1], *Lourdes Aguilar*[2], Juanma Garrido[3]

[1]Dpt. Computer Science, Universidad de Valladolid, Spain
[2]Dpt. of Spanish Philology, Universidad Autónoma de Barcelona, Spain
[3]Dpt. of Translation and Language Sciences, Universidad Pompeu Fabra, Spain
descuder@infor.uva.es, juamaria.garrido@upf.edu, lourdes.aguilar@uab.es

Entrainment in speech is commonly defined as a speaker's adaptation to the speech of his interlocutor [1]. This paper presents a corpus that contains a collection of dialogs that simulate real phone services. The preliminary tests that have been done on the prosodic features of the dialog turns evidence that the entrainment phenomenon is clearly observed. We defend the future use of this corpus for analyzing and modeling this phenomenon in technological applications.[1]

The corpus comprises two distinct data-sets, a news subcorpus and a dialogue subcorpus, the latter containing either conversational or task-oriented speech. More than twenty five hours were recorded by twenty eight speakers per language (Catalan and Spanish). Among these speakers, sixteen were professional (four radio news announcers and four advertising actors). The entire material presented here has been transcribed, aligned with the acoustic signal and prosodically annotated. All material contained in the corpus is provided under a Creative Commons Attribution 3.0 Unported License. This poster focuses on the task-oriented dialogs subcorpus and its application on the analysis of entrainment in conversational speech.

Our aim in the collection of this subcorpus is to offer a set of recorded interactions between two speakers oriented towards obtaining a specific goal. In all cases, speakers request for information, though for three different purposes: a) for traveling, b) for an exchange-university course, and c) for a tourist route.

These scenarios were selected following the diverse interests of research in areas such as (a) the development of speech technology dialogue systems (the design of automatic travel information systems, machine learning systems and tourist guides) and (b) linguistic studies that investigate the effect of variations derived from different communicative conditions on the speech of a given speaker.

The speakers who participate in each dialogue solve their task according to strict protocols, about which they were informed prior to their recording session, so that they would understand clearly what was expected from them (the protocols can be found in the technical report [2]. A relationship of cooperation is established between the participants, since both speaker and listener are involved in the execution of the task, and they both want to complete it with the maximum possible communicative success. These recordings stand as samples of intentional speech, similar to other kinds of intentional speech found in natural contexts, but obtained in a laboratory environment.

Due to our research interests on speech technology, all conversations were simulated to take place on the phone. For each conversation, one of the speakers plays the role of instruction-giver and the other, the role of instruction-follower. In order to avoid long silences or unnatural hesitations, both participants were provided with the information necessary to solve each task, and it was made sure that they would read it, and become familiar with each scenario before the recording started.

Travel information is the most formal task, since the scenario consists in a telephone conversation between an operator and a customer who requests for price information and time schedules of a specific route. A graph facilitated to the instructions-given to solve the task (see [3] for an explanation of the methodology).

Information request for an exchange-university course. This dialogue takes place between a staff member of a university administration office who provides information about a course at a foreign university, and a student who requests for it.

The information request for a tourist route is a type of interaction inspired by the Map Task [4]). Nevertheless, the description of the situation and the type of task are different. In the Map Task corpus, subjects are required to cooperate in order to reproduce on the follower's map the route printed on the giver's map, and the success of the communication is quantified by the degree of coincidence of both routes. In our corpus, however, one of the speakers plays the role of somebody who is planning a trip to the Greek island of Corfu, and telephones a colleague who has lived for five years in Greece, in order to request for specific information concerning the route on the island. There is no specific route to reproduce; there is only an initial and a final point of the trip, and some places to visit on the way.

Finally we recorded 72 task oriented dialogues which durations goes from 4 to 16 minutes. 24 of the informants recorded also free-speech dialogues. 4 of them participated also in the read aloud news subcorpus. The recording was done in studio conditions simulating the phone interaction. All the material has been transcribed, phonetically aligned and enriched with prosodic information.

[1] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels," in *Annual Meeting of the Association for Computational Linguistics (ACL/HLT)*, 2011, pp. 113–117.

[2] D. Escudero, C. V. L. A. Valentín Cardeñoso, C. de la Mota, J. Garrido, O. Larrea, and E. Rodero, "Proyecto glissando: Grabación de corpus prosódico de noticias y diálogos en español," Departamento de Informática, Universidad de Valladolid, Tech. Rep. IT-DI-2010-3, Decembre 2010.

[3] M. F. McTear, "Spoken dialogue technology: enabling the conversational user interface," *ACM Comput. Surv.*, vol. 34, pp. 90–169, March 2002.

[4] A. Anderson, M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert, "The hcrc map task corpus," *Language and Speech*, no. 34, pp. 351–366, 1991.

[1]This research has been funded by three research grants awarded by the Spanish Ministerio de Ciencia e Innovación, namely the *Glissando project* FFI2008-04982-C003-01,2,3 FFI2011-29559-C02-01,2

Do speakers make use of the visual channel to improve their intelligibility in adverse conditions? A pilot study.

Maëva Garnier[1], *Lucie Ménard*[2], *Gabrielle Richard*[2]

[1] Speech and Cognition Department, GIPSA-Lab, UMR CNRS 5216 & Grenoble Universités, France
[2] Laboratoire de phonétique, Université du Québec à Montréal, Canada
maeva.garnier@gipsa-lab.grenoble-inp.fr, menard.lucie@uqam.ca

1. Introduction

On one hand, it is now well known that seeing speech improves its perception, especially when speech is degraded by a noisy background [1]. On the other hand, some studies have shown that speakers adapt their speech production in noisy conditions. This adaptation, also called the « Lombard effect », mainly consists in talking louder and at higher pitch [2]. However, it is also accompanied by other speech modifications, such as increased amplitude and speed of lip articulation [3]. This raises the question of whether this hyper-articulation observed in Lombard speech can be considered as a communicative strategy to improve visual intelligibility.

This study aims at bringing elements of answers, by examining wether, in noise:

- speakers enhance significantly more their visible articulatory movements when their speech partner can see them compared to when the partner can only hear them.
- all the articulatory movements are enhanced similarly, or if the most visible ones (lips, jaw) are more enhanced than the others (tongue).

2. Material and Method

A French Canadian speaker was recorded while speaking in a quiet environment and in a cocktail-party noise of 85 dB played over loudspeakers. Three conditions of interaction were examined: (S1) No Interaction: The speaker read sentences aloud. (S2) Audio Only (AO): The speaker gave instructions to the experimenter who was standing at a writing board placed 2m in front of him and the back to him. (S3) Audio Visual (AV): The experimenter was standing at the same place as before, this time facing the speaker. Seven target-words were selected: /pap/, /pip/, /pup/, /pep/, /map/, /tap/, /nap/. They were produced in the carrying sentence « le mot ___ me plaît » (I like the word ___) and repeated ten times in each condition. In the two interactive conditions (S2 and S3), the speaker chose freely the order of production of the 70 sentences, so that the experimenter could not predict the target-word. The audio signal was recorded synchronously with the 3D movements of the lips, the jaw and the tongue, using electromagnetic articulography (Carstens AG 200).

3. Discussion

The results confirmed that all the speech modifications from a quiet to a noisy situation (increase of voice intensity and fundamental frequency, amplified movements of the lips and the tongue) are significantly greater when the speaker interact with a speech partner (S2 and S3), compared to when he only reads sentences aloud (S1).

As expected, acoustic modifications measured from a quiet to a noisy situation (increase of voice intensity, fundamental frequency and first formant frequency) were found to be significantly greater in a condition of AO interaction relative to a condition of AV interaction.

However, contrary to our hypothesis, articulatory modifications were not found to be greater in AV interaction. Very visible movements such as lip aperture, spreading, closure and protrusion, and jaw aperture had their amplitude and speed enhanced in noise. Nevertheless, this enhancement was significantly greater in AO interaction than in AV interaction. Less visible tongue movements were also significantly modified in noise, but only in the situation of AO interaction. For all vowels, the tongue was lower and more forward in the noisy condition compared to the quiet one. Tongue displacements were significantly amplified, with increased speed for the tip of the tongue and reduced speed at its root.

4. Conclusions

The results obtained from this speaker do not support the hypothesis that speakers modulate their production of visible cues in adaptation to the perceptual modalities of interaction. Instead, these results support the idea that all articulatory movements, regardless of their visibility, are enhanced similarly when speaking in noisy conditions and that this enhancement is primarily related to the increase of intensity. To compensate for the perturbation of intelligibility – which is greater in AO interaction than in AV interaction–, increasing voice intensity appears to be the main strategy. As a finer strategy, speakers do not seem to play on the visual channel to improve their intelligibility. The investigation of five other speakers will enable us to determine if these results can be widespread.

5. References

[1] Sumby, H. and Pollack, I. W., "Visual Contribution toSpeech Intelligibility in Noise." Journal of the Acoustic Society of America 26: 212-215, 1954.

[2] Junqua, J., "The lombard reflex and its role on human listener and automatic speech recognizers." Journal of the Acoustic Society of America 93(1): 510-524, 1993.

[3] Garnier, M., Henrich, N. and Dubois, D., "Influence of Sound Immersion and Communicative Interaction on the Lombard Effect." Journal of Speech, Language and Hearing Research 53(3): 588-608, 2010.

Priming, Timing, and the Phatic Component in Machine-Mediated Dialogue

Emer Gilmartin, Céline De Looze, Nick Campbell

Speech Communication Lab, Trinity College Dublin, Ireland

gilmare@tcd.ie, deloozec@tcd.ie, nick@tcd.ie

Conversation, as a human activity, meets an ecological need. Some of its elements communicate propositional content in order to perform a task, while others cement social bonds: forming the soundtrack to co-presence through friendly 'chat'. Early dialogue systems relied on information transfer to perform tasks, while newer systems attempt more human-like interaction [1]. A system that handles both chat and task-based dialogue would more closely model human-human communication and improve human-machine interaction. We present a data collection paradigm to capture chat and task-based elements in human-machine interaction. We first contrast task based and chat dialogue and then describe the collection and ongoing analysis of data.

1. Chat and Task-based Dialogue

In our view, dialogue contains dynamically varying proportions of chat and task-based elements. Chat and task-based dialogue may be contrasted in several areas:

Goal In chat, the goal is to establish co-presence, service interpersonal bonds, and maintain channels of communication. The goal may extend far beyond the dialogue duration and participants may be unaware the goal of their chat. In task-based dialogue, goals are the accomplishment of the task, and are more discrete, short term and clearly defined.

Content In task-based exchanges, where information transfer is vital, content may be positive, negative or controversial – it may be unwelcome to the receiver, but accomplishing the task overrules social considerations. Chat is light, without controversial content, and often phatic.

Structure Task based dialogue consists of often nested adjacency pairs, each of which performs a task, and which may be embedded to several levels as participants collaborate to build common ground. Chat, in our view, uses the same building blocks to provide a 'soundtrack' to interaction or co-presence, but its structure is shallower as deep embedding would cause a switch to task based dialogue.

Cognitive processing Much conversational analysis is based on a left-context or reactive view of dialogue where interlocutors respond to information received and use it to formulate their next contribution. While this left-context may hold true for task-based dialogue, chat relies more on feed forward control or prediction, where the speaker does not need to consciously analyse the content of the previous utterance, and may produce utterances at appropriate intervals based on past experience. Conversation fits a dual-process cognitive model, where a phenomenon can result both from an implicit unconscious process (commonly known as System 1) and from an explicit conscious one (System 2) [2]. Processing of chat can be shallow and fast, as per System 1, while creating a response to content critical information in a task-based dialogue is analagous to System 2.

2. WOZ Design and Data Collection

We designed a Wizard of Oz study to engage people in chat and/or task-based dialogues with a machine, using TCD Speech Communications Lab's robot platform, HERME [3] to engage visitors to an exhibition at the Science Gallery, TCD in a friendly chat with embedded task-based elements.

The dialogue was designed with reference to previous work on creating realistic human-machine dialogue [4]. The set script included a series of independent utterance sequences or 'volleys', of the form *statement/question - wait for response - feedback - wait - feedback*. This structure allowed the robot to control the dialogue as each sequence ending set the scene for the robot to start the next sequence. It also aided control by priming responses and confining the interlocutor to a limited range of responses. The number of utterances in each volley was limited, as the robot's control diminished with number of utterances or levels of embedding. These chat sequences were interspersed with task-based sections where the robot asked the user to sign a consent form, and distractors such as "I like your hair."

Over a three-month period in 2011, the robot was exhibited in the Science Gallery at TCD, engaging in conversations with over 500 participants, who were not paid subjects but random visitors who walked in off the street. When a prospective participant approached, face recognition software triggered the dialogue module, controlled remotely by a wizard observing over a Skype connection. The wizard pressed a key to step through the dialogue or to reset the conversation. The only freedom was in the timing of the key press, with the wizard instructed to initiate the next utterance when it seemed appropriate.

We are currently analysing the timing of gaps and overlaps and investigating the role of priming in the data, to better understand the characteristics of chat and task-based dialogue in order to inform the design of more human-like dialogue systems.

Acknowledgements
This work was undertaken as part of the FASTNET project - Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) 09/IN.1/I2631

3. References

[1] J. Edlund, J. Gustafson, M. Heldner, and A. Hjalmarsson, "Towards human-like spoken dialogue systems," *Speech Communication*, vol. 50, no. 8, pp. 630–645, 2008.

[2] D. Kahneman, *Thinking, Fast and Slow*. Farrar Straus & Giroux, 2011.

[3] F. Han, E. Gilmartin, C. De Looze, B. Vaughan, and N. Campbell, "The Herme database of spontaneous multimodal human-robot dialogues," in *Proc. LREC, Istanbul, Turkey*, 2012.

[4] T. Bickmore and J. Cassell, "'How about this weather?'-Social dialogue with embodied conversational agents," in *Proceedings of the AAAI Fall Symposium on Socially Intelligent Agents. North Falmouth, MA*, 2000.

Unsupervised Normal-to-Lombard Spectral Envelope Transformation; Examining Loudness, Voicing & Stationarity

Elizabeth Godoy[1], *Yannis Stylianou*[1], *Julián Villegas*[2]

[1]Institute of Computer Science, FORTH, Crete, Greece
[2]Ikerbasque, Language and Speech Laboratory, Universidad del Pais Vasco, Spain.

Abstract

When speaking in noisy environments, humans modify their speech in order to make it more intelligible: this phenomenon is known as the Lombard effect [1],[2]. It has been shown that, among the various Lombard modifications, those to the spectral envelope account for the largest increases in speech intelligibility [3]. The present work examines and seeks to exploit the spectral envelope differences between Normal and Lombard speech for multiple (4 male, 4 female) speakers of the GRID corpus in an unsupervised context, i.e., in the absence of segmentation or phonetic labeling. Our goals are twofold: 1) to transform the Normal speech spectral envelope towards that of the Lombard; 2) to isolate acoustic criteria that help to identify and better understand important spectral differences between Normal and Lombard speech.[1]

1. Analyses

The present analyses examine the "true" spectral envelopes (cepstral order 48) [4] of Normal and Lombard (speakers listened through headphones to 96dB speech-shaped noise) speech. In particular, we define a spectral envelope "correction" filter as the difference (in dB, with no DC component) between the average Lombard and Normal envelopes, calculated across all frames (pitch asynchronous analysis, 30ms Hanning window, 10ms step) for a given speaker. Essentially, these correction filters are able to clearly show regions in frequency for which the Lombard speech has more or less energy, as well as indicate overall Normal-Lombard differences in spectral envelope shape. Moreover, similarly to the idea behind the LPC-based modifications in [3], the correction filters can be applied to the magnitude spectrum of Normal speech for transformation towards Lombard speech. This transformation is inspired by the "amplitude scaling" proposed for voice conversion in [5].

In addition to differences in the spectral envelopes calculated across all frames, the present work examines two acoustic indicators that can be used to generate classes in an unsupervised context. One indicator is for voicing (i.e., the number of zero-crossings in a frame, normalized by the frame length) and the other is for stationarity (i.e., the "transition rate" of spectral events in a temporal decomposition of speech [6]).

Moreover, in order to get an idea about the perceived differences between the Normal, Lombard and transformed sentences, a loudness measure defined by the PEAQ standard is used [7]. In particular, a metric defined as the average difference in loudness between the Lombard and Normal speech in an inclusive formant region (500-4500Hz) is examined as a function of the acoustic criteria on voicing and stationarity.

2. Results

Examining the spectral envelope correction filters for each speaker, an overall trend is a boosting of energy in the 500-4500Hz region, confirming general observations on the Lombard effect [3]. The shape within this range can vary, with prominent maxima generally between 750-2250Hz and 3000-3500Hz for males and 1250-2750Hz and 3500-4500Hz for females, exhibiting differences due to vocal tract length.

[1]**Acknowledgements**: This work was supported by LISTA. The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 25623.

When used for Normal-to-Lombard transformation for each speaker, initial (informal) listening tests of the energy-normalized speech samples in noise (babble and white) suggest that the correction filters are generally effective at increasing intelligibility. This is similar to the results reported in [8], suggesting that estimating and applying spectral modifications on a more global (sentence or even corpus) level, compared to the frame-level, can significantly improve intelligibility.

In terms of the generalized acoustic classes, first, the spectral envelope correction filters for unvoiced frames demonstrate opposite trends from the voiced case (i.e., there is a reduction in energy in the mid-frequency 1-5kHz region); this is not surprising when considering the opposite negative/positive spectral slopes of voiced/unvoiced speech. However, the unvoiced frames appear to have little effect on loudness; in particular, the loudness metric indicates that the Normal speech is even slightly louder for unvoiced frames in the 500-4500Hz region. Unlike voicing, for the stationarity measure, there was little observed difference in the average spectral envelopes (both Normal and Lombard) between the stationary and transient parts of speech. However, it was noted that peaks in the transition rate indicate changes in loudness. Thus, the stationarity measure acts as a type of acoustic segmentation.

3. Perspectives

Observations from the present work indicate that the Lombard effect on the spectral envelope is not constant across speech; varying in time and frequency as a function of different acoustic events. Future work will try to isolate spectro-temporal regions in which the spectral envelope differences are most significant. The goal is then to have deeper acoustic understanding of Lombard effect and to define more effective transformation in unsupervised contexts.

4. References

[1] E. Lombard, "Le signe de l'elevation de la voix, annals maladiers oreille," *Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[2] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech producton: Acoustical and perceptual analyses," *J. Acous. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[3] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *SpeechComm*, no. 51, pp. 1253–1262, 2009.

[4] A. Roebel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Digital Audio Effects (DAFx)*, 2005, pp. 30–35.

[5] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans Audio, Speech, Lang Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[6] D. Kapilow, Y. Stylianou, and J. Schroeter, "Detection of non-stationarity in speech signals and its application to time-scaling," in *Eurospeech*, 1999, pp. 2307–2310.

[7] "ITU standard rec-bs.1387-1-2001," 2001.

[8] J. Villegas, M. Cooke, and C. Mayo, "The influence of temporal and spectral modifications on the intelligibility of normal and lombard speech," in *Proc. SPiN–2012: The 4 Int. Wkshp. on Speech in Noise: Intelligibility and Quality*, Cardiff, Jan 2012.

The Whispering Talker: Production and Perception of French Boundary Tones

Willemijn Heeren and Christian Lorenzi

Equipe Audition (CNRS, UMR LPP 8158), Ecole normale supérieure
Paris Sciences & Lettres, Paris, France
{willemijn.heeren;lorenzi}@ens.fr

1. Introduction

When a speaker wants to communicate under silent circumstances (e.g., in a library), he/she reverts to whisper. Intrinsic to whisper is the speaker's production of a noise source instead of the semi-periodic one resulting in voiced speech. As a consequence of this, speakers are expected to adapt their message acoustically, especially when expressing intonation, in order to be understood. We investigated how whispering speakers adapt to their listeners when expressing a prosodic contrast that is normally thought to be largely carried by the fundamental frequency, F0.

It has been well established that listeners have pitch percepts in whisper, [1–3]. The question how these percepts can be explained has been asked in only a handful of studies, many of which are impressionistic due to their age. So far, pitch percepts in whisper have been attributed to changes in formant frequencies, e.g. [1–4], but differences in intensity [2][4], and in some cases duration [1], have also been reported. But exactly which formants are involved, and are there other acoustic correlates that contribute to expressing pitch in whisper? And how relevant are these correlates for listeners?

To answer the research questions whether and how whispering speakers adapt their output to transferring intonation, parallel production and perception data were collected of normal and whispered speech. We have taken the case of boundary tones, signaling either an interrogative (H%) or an affirmative (L%). Perception was explored by dividing the speech spectrum into a number of frequency regions, and studying the relative information content of those regions to pitch perception as well as which acoustic cues could carry this information. Various vowel contexts were included to explore possible variation in expressing whispered intonation as well as restrictions set by the intonation-bearing unit.

2. Method

2.1. Production data collection

Twenty-five French determiner-noun combinations recorded with a high (H%) and with a low (L%) boundary tone. The final syllable of each noun contained one of five vowels (/i, a, e, u, o/), with five different words per vowel. A male and a female native speaker of French recorded the materials. Affirmative and interrogative targets were presented one by one on a computer screen, in a pseudo-random order. Live feedback was provided by a naïve listener sitting outside the booth, performing a classification task on each of the speaker's utterances, i.e. indicating if the utterance was perceived as interrogative or affirmative. Before the next target was presented, the speaker got feedback about the listener's understanding of the previous one. By keeping the listener outside the recording booth, and invisible to the speaker, the only cues the speaker could provide were auditory.

2.2. Perception experiment

A 2IFC discrimination task with feedback was run in which listeners identified the interval containing the interrogative. This was repeated in four different conditions: a broadband condition (50-8000 Hz) and three filter conditions (low-pass, band-pass, high-pass). Auditory frequency resolution (the auditory system's ability to hear out each spectral component composing a complex sound) varied across these three regions. Spectral components were expected to be resolved in the low-pass condition (<6th harmonic), and totally unresolved in the high-pass condition (>11th harmonic). The band-pass condition corresponded to an intermediate range (6th–11th harmonic), where spectral components were presumably partially resolved. These filtering conditions were used to assess the contribution of spectral information per region, and were constant across vowel contexts. Twenty normal-hearing adults participated (informed consent obtained).

3. Analysis/Results

Perception results were analyzed using repeated measures ANOVAs on the RAU scores, with within-subjects factors Filter Condition (4), Speaker Gender (2) and Vowel Quality (5). This revealed, amongst others, an interaction of Filter Condition by Vowel Quality, $F(8.8,166.4)=8.6$, $p<.001$, and main effects of Filter Condition, $F(2.3,43.6)= 69.0$, $p<.001$, and of Vowel Quality, $F(4,76)=14.6$, $p<.001$. Listener performance differed between filter conditions, with best performance in the broadband condition (80% correct), followed by the high-pass condition (70% correct), and the low-pass and band-pass conditions (62% and 54% correct). In addition, overall performance differed between vowels, and performance also varied between vowels within a condition.

All stimuli were passed through auditory models. Results confirmed that periodicity-pitch cues were absent, and that cues must be spectral in nature. Listener performance significantly correlated ($p<.01$) with excitation pattern-derived measures, such as slope, level of the highest peak against the background, and difference in centre of gravity between high and low versions of the same vowel. Variation in formant locations may explain performance differences between vowels (also within filtering conditions). Performance in the high-pass condition suggests this goes beyond F1 and F2.

4. References

[1] Fónagy, J., "Accent et intonation dans la parole chuchotée", Phonetica, 20:177-192, 1969.
[2] Heeren, W. and van Heuven, V. J., "Perception and production of boundary tones in whispered Dutch", Proc. Interspeech 2009, Brighton, 2411-2414, 2009.
[3] Higashikawa, M., Nakai, K., Sakakura, A. and Takahashi, H., "Perceived pitch of whispered vowels-relationship with formant frequencies: a preliminary study", J. Voice, 2:155-158, 1996.
[4] Meyer-Eppler, W., "Realization of prosodic features in whispered speech", J. Acoust. Soc. Am., 19:104-106, 1957.

A comparison of the effects of alteration to auditory feedback and speech motor learning

Peter Howell

Division of Psychology and Language Sciences, University College London
p.howell@ucl.ac.uk

Abstract

This poster argues that a distinction needs to be made between the effects of immediate auditory feedback and motor learning. The effects of all forms of altered feedback (delayed, changed in intensity or frequency) are immediate and transient whereas those involved in motor learning take longer to establish and can lead to sustained changes in speech production. Furthermore, there are some inherent features of auditory feedback which suggest it cannot be used to compensate for auditory changes in the short term. This includes speech that takes place without audition (hearing impaired), time required for auditory processing and so on. Work is reported which shows that alterations to auditory feedback affect low level, pre-linguistic systems in the cerebellum. For example, susceptibility of speakers to the effects of delayed auditory feedback correlates with performance in cerebellar tasks.

Whilst auditory feedback does not appear to be necessary for ongoing control of speech, there is some evidence that speakers compensate when changes are made to auditory feedback over longer terms. This suggests linguistic or articulatory information can be extracted from the feedback and used to modify the stored motor program for future productions (motor learning). This is a feature of the feedforward control system in the DIVA model [1]. However, recent work by Shiller and colleagues suggests that the process involved in motor learning is different from a classic feedback process. In particular, a feedback process requires a stable perceptual referent to guide production whereas [2] showed that when production and perception were biased by presenting sounds for categorization with unequal frequencies, perception and production readjusted. Some recent work using Shiller et al.'s paradigm with a novel speech contrast are reported.

References

[1] F. Guenther, "Neural modeling of speech production," in *Speech motor control in normal and disordered speech*, B. Maassen, W. Hulstjin, R. Kent, H. F. M. Peters, and P. H. M. M. van Lieshout, Eds. Nijmegen: Uttgeverij Vantilt, 2001, pp. 12–15.

[2] D. Shiller, M. Sato, V. Gracco, and S. Baum, "Perceptual recalibration of speech sounds following speech motor learning," *J. Acoust. Soc. Am.*, vol. 125, no. 2, pp. 1103–1113, 2009.

Characterizing listeners' performance in a speaking-while-listening task

Nandini Iyer[1], John Stewart[2], Sarah Sullivan[2], Douglas Brungart[3], Brian Simpson[1]
[1]Battlespace Acoustics Branch, Air Force Research Laboratory, WPAFB, OH 45433
[2]Ball Aerospace, Dayton OH 45433
[3]Walter Reed National Military Medical Center, Bethesda, MD 20889
Nandini.Iyer@wpafb.af.mil

Abstract

Communications in the operational world of the Air Force is extremely challenging. Operators listen to several channels of ongoing radio communication while responding to some or all of these channels, often performing several additional secondary tasks. And while much is known about the masking of an incoming speech signal by noise or speech, very little is known about the effectiveness with which listeners can perform in simultaneous "speaking while listening" situations; perhaps even more importantly, it is unclear in such situations how they might accomplish complex-decision making tasks. In his seminal work, [1] showed that significant interference occurred while performing such dual tasks; however it was not clear if the decrement in performance was because 1) listeners were limited in their ability to simultaneously process their own speech and the speech of others due to overlap in the auditory periphery between the incoming message and responses, or 2) deficits in performance were due to competition of resources because listeners were attending to what they had just said or were about to say while processing an incoming message. A speaking - while - listening task was adapted [1] to investigate the measure the deficits in such situations. In the task, listeners responded to a series of yes-no queries posed by a talker regarding a visual display on a computer monitor. Accuracy of the responses were scored based on three separate components of the message: 1) assigning the response back to the call-sign of the talker making the inquiry; 2) responding to the assigned listener call-sign and ignoring messages addressed to a distracter call-sign; and 3) providing the appropriate yes or no response to the inquiry. Two main variables were manipulated in the experiment: listener-talker gender configuration and rate of incoming messages. It was hypothesized that, if deficits in the task were related to energetic masking cased by the listener's spoken responses masking the incoming speech signals, then performance in the task would be worse in the case where the listener heard a same-gender talker with an acoustically similar voice. Alternatively, if deficits in the task were related to a competition of resources, then altering the rate of incoming message would correlate directly to the rate of missed messages. The results indicate that changing the rate of incoming messages had the largest impact on accuracy of listeners' responses to incoming messages. Additionally, performance deficits in the task could be counteracted by shortening the syntactic structure of the response phrase, suggesting that limitations in speaking-while-listening tasks may be largely due to an interference with response formulation while listening to an incoming message.

References

[1] Broadbent, D. E. (1952). Speaking and listening simultaneously. *Journal of Experimental Psychology*, **43(4)**, 267-273.

Speaking in quiet and in noise: Do auditory and articulatory properties pattern together?

Jeesun Kim, Chris Davis

MARCS Institute, University of Western Sydney

j.kim@uws.edu.au, Chris.Davis@uws.edu.au

1. Introduction

It is common that noise is present when speaking and listening and speech perception is detrimentally affected by it. When speaking in noise, talkers change the way they articulate and the way that speech sounds. In terms of acoustic properties, speech in noise typically has greater duration, an increase in F0 and increased energy at higher frequencies [1] and such speech is more intelligible than speech produced in quiet (when noise is mixed with both at the same signal to noise ratio, see [2]). Although the changes in acoustic properties that differentiate speech produced in noise from that produced in quiet have been studied at length, relatively little work has been done on the corresponding articulatory properties. It has been reported that speech in noise is produced with increased motion (both for movements directly related to articulation as well as those for that typically accompany speech, e.g., head motion [3]. Further, it has been shown that when a listener can see the talker, articulations produced in noise provide a more effective boost to intelligibility than those produced in quiet [4]. The aim of the current research is to provide some basic information about the way that auditory and articulatory properties change as a function of the production environment (in quiet or in noise) and the relationship between these changes. This was done by measuring auditory and articulatory (via motion tracking) properties of speech produced in quiet and noise and leveraging the variability across the productions of 8 speakers.

1.1. Method

Eight talkers participated in the speech data capture session. All were native speakers of English. The spoken materials were 10 sentences selected from the Harvard list of phonetically balanced sentences (IEEE 1969). The background noise consisted of a babble track of three female and one male talker. Twenty-eight movement sensors were used to measure articulation, 24 placed on the face (jaw, lips, cheeks, brows) and four rigid body markers on a head-rig. These sensors were tracked with a 3020 Optotrak system. The talker spoke each sentence in quiet and in noise. In the noise condition, the talker wore a set of earphones and heard the babble track as they spoke each sentence (this was played at approximately 80 dB SPL, similar to the level used by [2]). Two tokens of each sentence were recorded in each condition. The auditory data was recorded using one of two lapel-worn microphones (Shure SM12A; Senheiser e840). The recordings were analysed using Praat (Version 4.4.16, [5]) with consonant and vowel segments labelled. The articulatory motion data was reduced by a data-driven approach to guiding the PCA decomposition. In this guided PCA approach a subset of markers was defined so that the analysis could be conducted in terms of interpretable components, with components derived using these selected motion data by successive subtraction once rigid head movements have been removed (see [6]; [7]). The set of markers used to guide the PCA were selected to reflect articulatory components: gPCA 1 for Jaw motion in the Y axis; gPCA 2 for Mouth Opening in the Y axis; gPCA 3 for Lower Lip motion in the Y axis; gPCA 4 for Upper Lip motion in the Y axis; gPCA 5 for Lip motion (spreading) in the X, Y, and Z axes; and gPCA 6 for Jaw Protrusion in the Z axis.

1.2. Results and Conclusion

The acoustic analysis showed that, collapsed over all talkers, speech produced in noise had a greater duration, a greater F0 and increased energy at higher frequencies. However, there were significant differences across participants for all of these measures. Analysis of the motion data showed that overall jaw motion (in the Y axis and protrusion) and mouth opening and lip rounding were significantly greater for in-noise than in-quiet productions. Once again there was considerable variation in these measures across participants. In regard to acoustic-articulatory relationships, there was a straightforward relationship between average intensity and average jaw motion. Talkers whose speech in noise showed the greatest difference in intensity (relative to their in-quiet speech) also had the largest relative jaw motion. However, other acoustic-articulatory relationships (particularly those reflecting changes measured over time) were more complex and variable.

2. References

[1] Lu, Y. & Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. Journal of the Acoustical Society of America, 124(5):3261-75.

[2] Junqua, J-C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. Journal of the Acoustical Society of America, 93: 510-524.

[3] Kim, J., Davis, C., Vignali, G., & Hill, H. (2005). A visual concomitant of the Lombard reflex. Proceedings of AVSP 2005, 17-22.

[4] Kim, J., Davis, C., & Sironic, A. (2011). Hearing speech in noise: Seeing a loud talker is better. Perception, 40: 853-862.

[5] Boersma P, & Weenink D. (2005). Praat, http://www.fon.hum.uva.nl/praat/

[6] Badin, P., Bailly, G., Reveret, L., Baciu, M., Segebarth, C. & Savariaux, C. (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. Journal of Phonetics, 30: 533-553.

[7] Maeda, S. (2005). Face models based on a guided PCA of motion capture data: Speaker dependant variability in /s/-/ʃ/ contrast production. ZAS Papers in Linguistics, 40: 95-108.

On the detection of the intelligibility advantage of clear speech vs. casual speech

M. Koutsogiannaki[1], *C. Mayo*[2], *V. Kandia*[1] *and Y. Stylianou*[1]

[1]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
[2] Centre for Speech Technology Research, the University of Edinburgh, UK
{mkoutsog, vkandia, yannis}@ics.forth.gr, catherin@ling.ed.ac.uk

1. Introduction

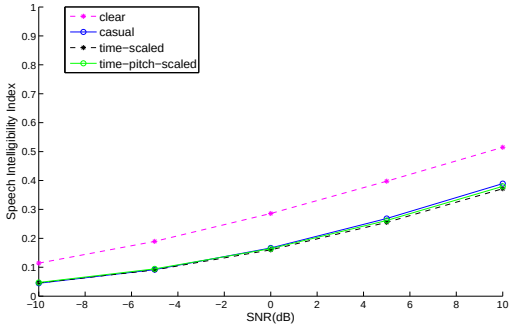
This work focuses on speaking rate and pitch differences between clear and casual signals. Transforming the clear signal to match the casual signal in terms of the prosodic features, the effect of each one factor to the intelligibility advantage of clear speech is examined based on acoustic evaluations and objective measures.

2. Methodology and results

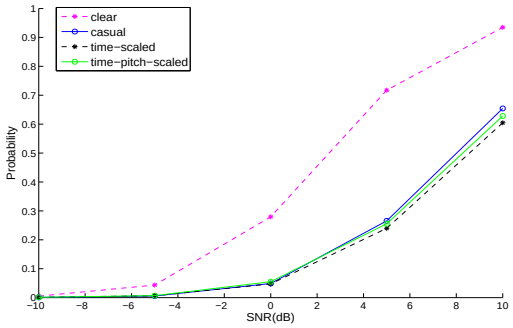
Studies show that even though intelligibility is increased by the decrease of speech rate both in clear and casual speech, clear speech can also be produced without the decrease of rate after training the speakers [1]. This suggests that clear speech has inherent acoustic properties independent of rate, that contribute to improved intelligibility. Many speakers in their effort to elicit clear speech change their pitch both in level and range. However, it is not luminous if pitch modification is a feature that contributes to intelligibility.

In this work, we examine if clear speech signals are still more comprehensive than casual speech signals after equalizing the prosody features on the two signals. To this purpose, a database of clear and casual speech signals is analyzed. Speakers in this database read sentences both in clear and casual way [2]. Clear speech sentences are modified in duration and pitch to match the corresponding attributes of casual speech signals. After the equalization, pilot acoustical test analysis and objective measure tests are performed on the four equal set of signals; on the initial database of clear and casual signals and additionally on the time-scaled and time and pitch-scaled clear signals.

In the acoustical pilot experiments, speech shaped noise is added to the signals to create the test signals, with Signal to Noise Ratio of $0dB$. Results show that on a set of pairs of clear and casual sentences, in 64% of the cases listeners found more intelligible the clear sentences. However, in time-scaled and time-pitch-scaled modified clear sentences intelligibility scores were deteriorated. Objective measure tests were also performed, using a modified version of the extended Speech Intelligibility Index (SII) [3]. SII was evaluated in a separate database giving high correlation scores with perceptual acoustical tests. According to the SSI measure, clear signals have higher intelligibility scores than casual signals (Fig.1(a)) with higher probability (Fig.1(b)) of identifying a sentence for SNR levels above $-5dB$. On the other hand, casual signals, time-scaled and time-pitch-scaled clear signals that have the same duration, give the same score of SII independent of the SNR level (Fig.1(a)). Pilot acoustical experiments and objective measures suggest that duration indeed plays a significant role to intelligibility, whereas pitch modifications do not seem to contribute to intelligibility.



(a)



(b)

Figure 1: Objective Measure Score for the four set of signals for different levels of SNR. a) Speech Intelligibility Index b) Probability of correctly identifying a sentence

3. References

[1] J. Krause and L. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *JASA*, vol. 115, no. 362-378, 2004.

[2] V. Hazan and R. Baker, "Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style?" *DiSS-LPSS*, pp. 7–10, 2010.

[3] K. S. Rherbergen and N. J. Versfeld, "Speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *JASA*, pp. 2181–2192, 2005.

Automaticity and consciousness in phonetic convergence

Natalie Lewandowski[1]

[1]Institute for Natural Language Processing, University of Stuttgart, Stuttgart

natalie.lewandowski@ims.uni-stuttgart.de

Abstract

The abstract describes a study of native-nonnative conversations, with a focus on an objective acoustic measurement of phonetic convergence. Phonetic convergence occurs when the segmental and suprasegmental pronunciation of two speakers becomes more alike [1]. Comparing amplitude envelopes between an early and late point in the dialogs revealed that despite the speakers' strong intention, total control of convergence mechanisms might in fact not be feasible.

Index Terms: phonetic convergence, accommodation, nonnative

1. Experiment

One highly debated issue in the field of convergence concerns the automaticity vs. social motivation of the accommodation of two interacting partners. The first strand of research opts for the existence of a rather automatic alignment process [2], while the latter highlights that the speakers can make the decision to converge or diverge consciously, based a.o. on the social context (Communication Accommodation Theory, [3]). There have also been voices for a mixed model of convergence, involving both possibilities – automatic and controlled behavior [4].

Twenty speakers of German were involved in two dialogs with English native speakers (J and T). While the German speakers were not informed about the motivation of the experiment, both English native speakers were aware that the goal was the investigation of phonetic convergence in native-nonnative interactions. Moreover, they were explicitly asked not to alter their pronunciation to avoid accommodating the nonnative speakers' accents. The dialogs lasted between 15 and 20 minutes and were elicited using a picture matching game – the Diapix [5].

2. Results and discussion

An analysis of the amplitude envelope signals [6] of manually extracted target words from both speakers (from the beginning and the end parts of the dialogs) revealed that the English native speakers converged to their German speaking partners. The two amplitude envelopes were compared via a cross-correlation which returned a *match value* of the two signals (the closer to 1, the more similar to each other the envelopes). Figure 1 displays the mean convergence in all 20 dialogs, as the difference between the late and early match values of the target words' amplitude envelopes ($p<.01$ for both native speakers). This positive shift in pronunciation toward the nonnative partners occurred not only in spite of the request for controlling their pronunciation but also without their conscious knowledge. Both participants reported after the task that they felt they had succeeded in maintaining their own way of speaking, which contradicts the objective acoustic measurements performed on the data.

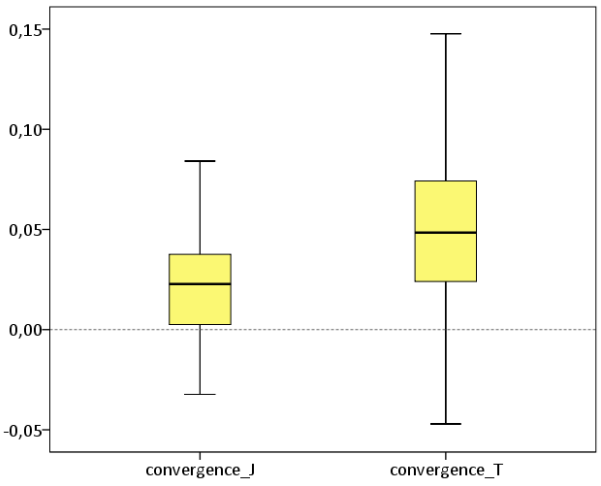


Figure 1: *Convergence of the two native speakers – J and T.*

The results speak for a hybrid model of convergence, which would indeed take into account automatic tendencies combined with other influencing factors, such as social and psychological features. This implies that the default tendency in communicative interaction seems to be convergence running largely subconsciously, which however, does not imply a total lack of control over the behavior. The observed variation in the native speaker data suggests the existence of other factors, mediating the degree of convergence, such as the partners' perceived level of proficiency in English, and other social and contextual factors determining the communicative situation.

3. References

[1] Pardo, J. S., "On phonetic convergence during conversational interaction", in J. Acoust. Soc. Am., 119:2382–2393, 2006.

[2] Pickering, M. J. and Garrod, M., "Toward a mechanistic psychology of dialogue", in Behav. Brain Sci., 27(2):169–190, 2004.

[3] Giles, H. and Ogay, T., "Communication Accommodation Theory", In B. B. Whaley and W. Samter [Ed], Explaining Communication: Contemporary theories and exemplars, 293–310, Mahwah, 2006.

[4] Krauss, R. M. and Pardo, J. S., "Commentary on Pickering and Garrod. Is Alignment always the Result of Automatic Priming?", in Behav. Brain Sci., 27(2):203–204, 2004.

[5] van Engen, K. J. et al., "The wildcat corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles", in Lang Speech, 53(4):510–540, 2010.

[6] Wade, T., Dogil, G., Schütze, H., Walsh, M. and Möbius, B., "Syllable frequency effects in a context-sensitive segment production model", in JPhon, 38:227239, 2010.

Vowel Creation by Articulatory Control in HMM-based Parametric Speech Synthesis

Zhen-Hua Ling[1], *Korin Richmond*[2], *Junichi Yamagishi*[2]

[1]iFLYTEK Speech Lab, University of Science and Technology of China, P.R.China
[2]CSTR, University of Edinburgh, United Kingdom

zhling@ustc.edu, korin@cstr.ed.ac.uk, jyamagis@inf.ed.ac.uk

Hidden Markov model (HMM)-based parametric speech synthesis has become a mainstream speech synthesis method in recent years. This method is able to synthesise highly intelligible and smooth speech sounds. In addition, it makes speech synthesis far more flexible compared to the conventional unit selection and waveform concatenation approach. Several adaptation and interpolation methods have been applied to control model parameters and so diversify the characteristics of the generated speech [1]. However, this flexibility relies upon data-driven machine learning algorithms and it is difficult to integrate phonetic knowledge into the system directly when corresponding training data is not available. In previous work, we have proposed a method to improve the flexibility of HMM-based parametric speech synthesis further by integrating articulatory features [2]. Here, we use "articulatory features" to refer to the continuous movements of a group of speech articulators, such as the tongue, jaw, lips and velum, recorded by human articulography techniques. In this method, a unified acoustic-articulatory HMM is trained. The dependency between acoustic and articulatory features is modelled by a group of linear transforms which are either trained and tied context-dependently [2] or switched in the articulatory feature space [3]. During synthesis, the characteristics of the synthetic speech can be controlled by modifying the generated articulatory features according to phonetic rules.

In this paper, we apply this method of articulatory control to the task of vowel creation in HMM-based parametric speech synthesis. In this task, the target vowel to be created does not occur in the training set, but its phonetic characteristics are known beforehand. We aim to produce this target vowel effectively at synthesis time once appropriate articulatory representations are provided. This is potentially useful for applications such as speech synthesis for limited resource languages, cross-language speaker adaptation, and so on. In our previous approach, articulatory features are treated as HMM observation vectors on which the acoustic features depend. In contrast, in this paper we treat the articulatory features as external explanatory variables for the mean vectors of Gaussians to make it simpler to control synthetic speech via articulation. This model is called a "multiple regression HMM" (MRHMM). Our feature-space transform tying strategy [3] is also applied here. Furthermore, we remove vowel identity from the set of context features used during context-dependent model training in order to ensure compatibility between the estimated model parameters and the articulatory features of a new target vowel at synthesis time.

This work is partially funded by the National Nature Science Foundation of China (Grant No. 60905010) and the National Natural Science Foundation of China - Royal Society of Edinburgh Joint Project (Grant No. 61111130120). The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 256230 (LISTA), and EPSRC grants EP/I027696/1 (Ultrax) and EP/J002526/1 (CAF).

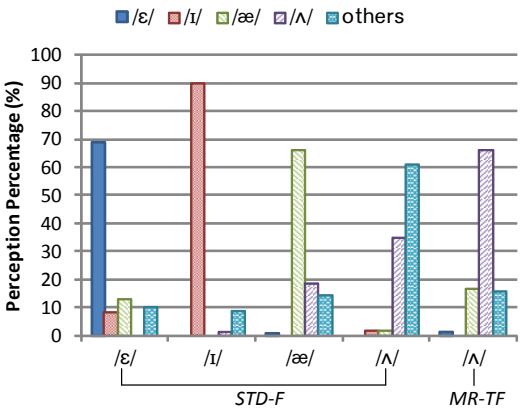


Figure 1: Vowel identity perception results for synthesising different vowels using the baseline system *STD-F* and creating vowel /ʌ/ by articulatory control using the proposed system *MR-TF*.

We have carried out a vowel identity perception test to evaluate the effectiveness of creating the target /ʌ/ vowel. Five monosyllabic words ("but", "hum", "puck", "tun", "dud") containing the /ʌ/ vowel were selected and embedded within a carrier sentence "Now we'll say ... again". For the purpose of comparison, we substituted the vowel /ʌ/ in the five monosyllabic words with /ɛ/, /ɪ/ and /æ/, and then synthesised the respective test sentences using the baseline system. Thirty-two native English listeners were asked to listen to these stimuli and to write down the key word in the carrier sentence they heard. These results are shown in Fig. 1. We see that only 35% of the synthesised vowels /ʌ/ were perceived correctly using the baseline system, due to the lack of acoustic training samples for this vowel. Using the proposed system and the generated articulatory features, this percentage increased to 66.25%, which is close to the perception accuracy of synthesising vowel /ɛ/ (68.75%) and /æ/ (66.25%) using the baseline system.

1. References

[1] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[2] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, Aug. 2009.

[3] Z.-H. Ling, K. Richmond, and J Yamagishi, "Feature-space transform tying in unified acoustic-articulatory modelling for articulatory control of HMM-based speech synthesis," in *Interspeech*, 2011, pp. 117–120.

The rate of intelligibility change with level for continuous speech

Alexandra MacPherson [1] *and Michael A Akeroyd,* [1]

[1]MRC Institute of Hearing Research, Glasgow Royal Infirmary, 16 Alexandra Parade, G31 2ER

alex@ihr.gla.ac.uk

1. Introduction

When listening to speech in noisy environments, increasing the level of the speech in comparison to that of the background noise usually increases its intelligibility. A listener may increase the volume on their TV or radio to better hear it, or a talker may raise their voice to be better understood. The amount of perceptual benefit that a listener will actually receive from this improvement in signal-to-noise ratio (SNR), however, is not fixed and instead depends entirely on the slope (gradient) of the psychometric function.

The psychometric function describes the relationship between the relative level of speech and its intelligibility. The shallower the slope of the psychometric function then the less benefit a listener will receive from any gain in speech level that may be offered. We have completed a systematic review of 913 psychometric functions which demonstrated that the median slope for masked speech is 6.4% per dB but that importantly some listening situations give shallower slopes than others. Reduced context, modulations and competing speech, for example, all act to flatten the slope.

All the studies identified in the review measured the slope of the psychometric function using short speech tokens: syllables, words, or sentences. Much of the speech we listen to on a daily basis, however, is not presented one word or sentence at a time; instead we are often required to keep up with a flow of information. The aim of the current study was, therefore, to develop a paradigm to measure the slope of psychometric functions for continuous speech.

2. Method

Seventeen listeners (mean age = 68, mean Hearing Loss = 38 dB HL) took part in the study. Participants listened to four-minute long extracts of an audio book over headphones while reading a printed transcript of the same text. Each transcript contained a set number of words which had been intentionally changed and therefore no longer matched those in the audio. The participant's task was to mark on the transcript any words that did not match those they had heard in the speech. The continuous speech extracts were played in a speech-shaped static noise and performance on the task was measured at seven different SNRs, selected individually for each listener. Psychometric functions were then constructed by calculating the percentages of mismatches correctly identified at each of the seven SNRs.

Seven-point psychometric functions were also measured for each listener using a standard speech-in-noise task. In the standard task, ASL sentences were presented one at a time in a speech-shaped static noise. The listener was asked to repeat each sentence and was given as much time as they needed to respond. Psychometric functions were constructed using the percentage of words correctly identified at each SNR. All psychometric functions were then fitted with a logistic function which was used to derive the slope at 50% correct.

3. Results

The mean slope for the continuous task (3.1% per dB) was found to be significantly shallower than the mean slope for the standard speech-in-noise task (11.0% per dB). This difference in slope is illustrated in Figure 1. The results suggest that listeners may receive considerably less perceptual benefit per decibel improvement in signal level in more realistic listening situations than standard tests would suggest.

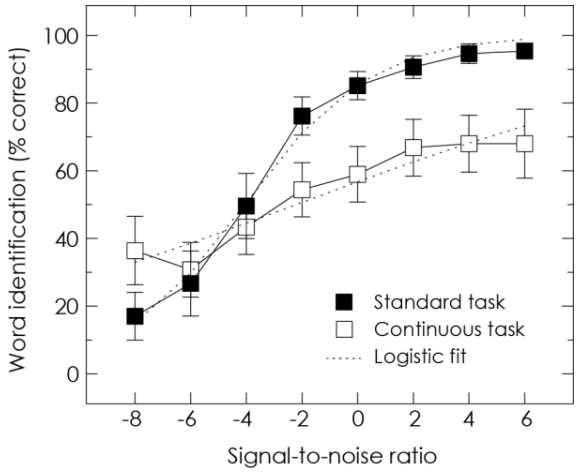


Figure 1: *Across listener mean psychometric functions measured in a continuous speech task and a standard speech-in-noise task.*

4. Acknowledgements

This work was supported by a MRC PhD studentship to the first author, the Medical Research Council (grant number U135097131), and by the Chief Scientist Office of the Scottish Government.

Effect of prosodic changes on speech intelligibility

Catherine Mayo[1] *& Vincent Aubanel*[2]

[1]Centre for Speech Technology Research, University of Edinburgh
catherin@inf.ed.ac.uk
[2]Language and Speech Laboratory, Ikerbasque and University of the Basque Country
v.aubanel@laslab.org

1. Abstract

Clear speech—that is, speech produced for listeners with perceptual and/or linguistic deficits (e.g., hearing impaired listeners, second language learners)—is characterised in the prosodic domain by a slower speech rate and by changes in fundamental frequency (F0) compared to conversational or plain speech (see review in [1]). Clear speech has also been found to be more intelligible than plain speech when both speech types are presented in noise. However, those studies that have specifically examined the role of speech rate and F0 in improved speech-in-noise intelligibility have found mixed results [2, 3, 4].

In the current study we examine multiple characteristics of speech rate and F0, and relate these to subjective and objective intelligibility-in-noise measurements, with the aim of more fully describing this relationship.

A male, native British English talker was recorded producing 25 randomly selected TIMIT sentences in five speech styles: plain, infant-directed, computer-directed, foreigner-directed and shouted. All but the shouted speech was elicited via instructions to the talker (e.g., "Please speak as if you were talking to a computer"). The shouted speech was produced while the talker listened to 90-95dB 9-speaker babble-shaped noise.

Human perceptual intelligibility-in-noise tests were conducted both in carefully monitored laboratory listening conditions, and using Amazon Mechanical Turk (AMT) crowdsourcing methods (noise: 9-speaker babble-shaped noise; signal-to-noise ratio: 0dB). Results from both sets of listeners indicate that this talker produced computer-directed speech that was significantly more intelligible than all other speech types. Additionally, all listeners found infant-directed speech to be less intelligible than all other speech types; for AMT listeners (who showed significantly higher word error rates than lab-based listeners) this difference in intelligibility was significant. An objective model of energetic masking based on glimpsing [5], which counts the extent of the spectro-temporal plane which escapes masking, accounts reasonably well for listener error rates.

Acoustic analysis of the recorded speech provides some possible explanations for these results. All non-plain speech styles displayed longer overall sentence durations than plain speech, with the exception of infant-directed speech; computer-directed speech showed the greatest overall durational increase. This increase in overall sentence duration is reflected in individual segment durations: when divided into sound classes, computer-directed speech had significantly longer segment durations than normal speech, the longest being vowels and nasal consonants. In contrast, infant-directed speech segment durations were statistically identical to normal speech. The durations of unfilled pauses were significantly longer for all non-plain speech styles than for plain speech. In addition, computer-directed speech showed a significant increase in number of unfilled pauses (+170%) over plain speech, as did foreigner-directed speech, though to a lesser extent.

Regarding pitch variation, all speech styles had a significant higher median F0 than plain speech, with the exception of computer-directed speech which was statistically identical. Furthermore, in computer-directed speech, F0 range was significantly lower than in plain speech. In contrast, F0 range in infant-directed speech was significantly higher than that in plain speech.

Taken together, acoustic analysis suggest that much of the prosody-related intelligibility gain comes from durational increases, and in particular from unfilled pauses, which are not accounted for by current models of objective intelligibility. Additionally, the least intelligible speech style, infant-directed speech, had the highest median pitch and widest pitch range, while the most intelligible speech style, computer-directed speech, had the lowest median pitch and flattest pitch range. These latter results contrast with earlier studies [4, 6], and suggest that the role of F0 in intelligibility may be more complex than previously posited.

2. References

[1] R. M. Uchanski, "Clear speech," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds., pp. 207–235. Blackwell, Oxford, 2005.

[2] V. Hazan and D. Markham, "Acoustic-phonetic correlates of talker intelligibility for adults and children," *Journal of the Acoustical Society of America*, vol. 116, pp. 3108–3118, 2004.

[3] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, pp. 1253–1262, 2009.

[4] J. S. Laures and K. Bunton, "Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions," *Journal of Communication Disorders*, vol. 36, pp. 449–464, 2003.

[5] M. Cooke, "A glimpsing model of speech perception in noise," *Journal of the Acoustical Society of America*, vol. 119, pp. 1562–1573, 2006.

[6] P. J. Watson and R. S. Schlauch, "The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours," *American Journal of Speech Language Pathology*, vol. 17, pp. 348–355, 2008.

An electropalatographic study of consonant production in Greek Lombard speech

Katerina Nicolaidis

Department of Theoretical and Applied Linguistics, Aristotle University of Thessaloniki, Greece

knicol@enl.auth.gr

Abstract

The present study uses the technique of electropalatography to examine lingual articulation during consonantal production in speech produced in quiet and noise. The results show that there are important spatio-temporal modifications in Lombard speech and presence of variability. In addition to changes potentially related to increased vocal effort there is some first evidence pointing towards speech modifications that can be communicatively driven.

Index Terms: Lombard speech, consonant articulation, electropalatography, duration, Greek

1. Introduction

Previous research has shown that speech production in noise adapts along several articulatory and acoustic parameters (e.g. [1], [2]). A central theoretical issue concerns the mechanisms that underlie these modifications, i.e. changes are due to a physiological automatic response, a communicative adaptation or a combination of both [3], [4], [2]. While acoustic parameters have been relatively well-studied there are limited physiological studies examining articulatory changes. In addition, studies on segmental production have focused on the acoustic changes of vowels leaving consonant production largely unexplored. This study analyses articulatory data on consonant production aiming to provide insights onto the nature of the modifications present.

2. Method

Electropalatography (British system, Articulate Instruments) was used to record four speakers (two male and two female) of Standard Modern Greek producing disyllabic words (C_1VC_2V) in the carrier phrase "leɣe ˈCVCV ˈpali" ('say___again'). C_2 was examined (/t, k, s, x, n, l, r/), C_1 was /p, b, f/, and V= /i, a/. Symmetrical sequences were recorded with stress on the first syllable, e.g. /ˈpasa/. Subjects produced five repetitions of the speech material in a quiet and in a noise condition. For the latter, multitalker babble noise, calibrated at 88dB SPL at the speaker's ears, was played over loudspeakers.

For the duration measurements, onset and end of consonantal constriction was marked. For the articulatory analysis, the first frame of maximum contact over the entire palate was annotated for /t, n, l, r/ and the first frame of maximum constriction for /s, x/. The temporal midpoint was annotated for the examination of coarticulatory effects. The total amount of contact in the alveolar and palatal regions, place of articulation for all consonants, degree of constriction for the fricatives, V-to-C coarticulatory effects and token-to-token variability were analysed using measures such as percentage frequency of electrode activation over repetitions, front and back totals, front and back Centre of Gravity, front and back mean lateral measure and variability index.

3. Results

Consonant production in Lombard speech was characterised by: (a) generally shorter durations for all consonants (with significant differences for /x, n/ only), (b) a tendency for less contact in the palatal region for /t, s, n, l, r/ suggesting a lowered tongue dorsum, (c) a tendency for more contact in the alveolar region for anterior consonants and significantly more anterior placement for /n, l, r/; for the latter consonants, more peripheral articulation and greater amount of contact may suggest stronger, hyperarticulated productions, (d) systematically smaller coarticulatory effects on the tongue dorsum and tip/blade for the anterior consonants, (e) no differences in amount of contact for /k, x/ in the two conditions. Variability to these patterns due to speaker, gender, consonant, vowel and condition was found. Increased speaker variability (c.f. [1]) may relate to speaker–dependent differences in consonant production and in adaptive behaviour in the presence of noise. Gender differences included greater consonantal duration for the female speakers, greater amount of contact and smaller coarticulatory effects possibly relating to the greater intelligibility of female speakers in Lombard speech as reported in previous studies (c.f. [1]). Contextual effects in the two conditions were in line with the model of articulatory constraints [5].

4. Conclusions

Although increased vocal effort may relate to some of the patterns found, there is some first evidence that speakers use adaptive strategies to compensate for the severe output constraints present in Lombard speech in line with the framework of adaptive variability proposed by the H&H model [6].

5. References

[1] Junqua, J-C. "The Lombard reflex and its role on human listeners and automatic speech recognisers", Journal of the Acoustical Society of America, 93(1): 510-524, 1993.

[2] Garnier, M., Henrich, N. and Dubois, D. "Influence of sound immersion and communicative interaction on the Lombard effect", Journal of Speech, Language and Hearing Research, 53: 588-608, 2010.

[3] Lombard, É. "Le signe de l'élévation de la voix", Ann. Maladies Oreille, Larynx, Nez, Pharynx, 37 : 101-119, 1911.

[4] Lane, H. and Tranel, B. "The Lombard sign and the role of hearing in speech", Journal of Speech and Hearing Research, 14: 677-709, 1971.

[5] Recasens D., Pallarès M. D. and Fontdevila, J. "A model of lingual coarticulation based on articulatory constraints", Journal of the Acoustical Society of America, 102(1): 544-561, 1997.

[6] Lindblom, B. "Explaining phonetic variation: a sketch of the H&H theory", in W. J. Hardcastle and A. Marchal [Eds] Speech Production and Speech Modelling, 403-439, Kluwer Academic Publishers, 1990.

Consonant production control in a computational model of hyper & hypo theory (C2H)

Mauro Nicolao, Roger K. Moore

Speech and Hearing Group, Dept. Computer Science, University of Sheffield, UK

m.nicolao@dcs.shef.ac.uk, r.k.moore@dcs.shef.ac.uk

According to speech production theories such as Lindblom's Hyper&Hypo theory [1], the goal of human speech communication is to transfer information from the talker to the listener minimising the effort and maximising the effectiveness. In order to achieve this, humans adjust their speech production according to the context in which they are speaking, paying particular attention to the listener's needs [2]. Energy distribution and organisation also play a crucial role in modelling these adjustments, especially to increase intelligibility in noise [3].

The experiment described here represents a natural extension our previous work [4], in which it is hypothesised that there are low-energy attractors in the human speech production system and, moreover, that an interpolation/extrapolation along the key dimension of hypo/hyper-articulation can be obtained by controlling the distance to such an attractor. Low-energy attractor refers to an acoustic configuration towards which human speech production tends to converge when hypo-articulated. The effort in the articulatory movements needed to pronounce different phones is reduced and a common acoustic realisation might be observed. This hypothesis was tested by reducing/increasing the acoustic distance between the standard vowels and the mid-central vowel produced by a HMM-based synthesiser with a scalable adaptation of the statistical models.

The same framework is used here to control consonant production and an evaluation of intelligibility for the transformed speech is shown. In the consonant domain, a unique low-energy attractor cannot be identified so easily as in the vowel domain, thus a more sophisticated concept of low-energy attractor is introduced. If the low-energy configuration of a phone is hypothesised to be an acoustic realisation which does not allow for a clear discrimination between it and its acoustically closest competitor, the hypo- and hyper-articulation configuration can be controlled by moving the acoustic characteristics of a consonant respectively towards or away from those of its competitor. Therefore, low-contrastive phone-pairs have been identified and the transformations to convert one into the other and vice versa have been trained. The result is a set of adaptations (using MLLR), which transform the standard pronunciation of a phone (e.g. [b]) towards its closest competitor (e.g. [v]) to achieve the hypo-articulated production, and in the opposite direction for the hyper-articulated one. This experiment is part of a wider project which aims to investigate the possibilities for introducing a feedback path into automatic speech synthesis such that adjustments can be made continuously as a function of perceived effectiveness in the communicative context. A computational model for H&H theory (C2H) has been developed using a modified HMM-based synthesiser and an auditory feedback system. The loop is needed to evaluate the speech intelligibility of the synthesiser outcome in the environmental disturbance and to control the strength of the adaptation. The modified synthesiser allows for recursive parameter generation; therefore it handles the auditory-loop information frame-by-frame and adapts the speech production dynamically. In this experiment, C2H is used to apply the consonant adaptation to a synthetic voice and to evaluate the outcome. The auditory feedback in the framework is obtained with Speech Intelligibility Index (SII) [5], a standard automatic index of intelligibility for speech in noise.

As in [4], a set of 200 utterances are synthesised with standard, fully hyper- and fully hypo-articulated pronunciation, and compared with the same intensity and same quality noise. The distribution of intelligibility differences, in terms SII, are summarised in Fig. 1 and the average values are $\Delta SII_{hyper-std} = 27.3 \pm 8.3\%$ and $\Delta SII_{hypo-std} = -8.3 \pm 5.2\%$.

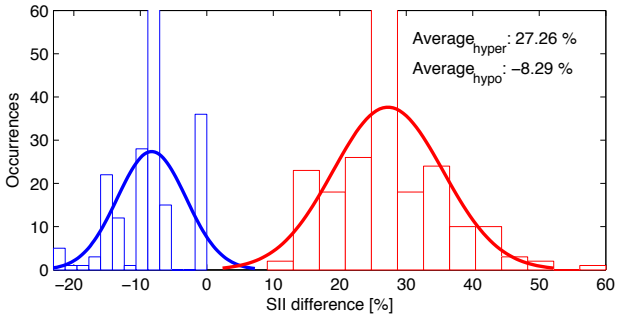


Figure 1: Distribution of SII differences between fully-hypo- (left-hand histogram) and fully-hyper-articulated (right-hand histogram) speech and standard HTS production.

It is concluded that the consonant production control tested in this experiment can be an effective tool to modify the intelligibility of synthetic speech in noisy environments. It is worth emphasising that such control emerges from phonetic-contrast alone.

This research was funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n 213850 - SCALE.

[1] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," *Speech production and speech modelling*, 1990.

[2] R. K. Moore, "PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction," *IEEE Transactions on Computers*, 2007.

[3] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," in *INTERSPEECH 2010*, 2010.

[4] R. K. Moore and M. Nicolao, "Reactive Speech Synthesis: Actively Managing Phonetic Contrast Along an H&H Continuum," in *ICPhS 2011*, 2011.

[5] ANSI, "American National Standard Methods for Calculation of the Speech Intelligibility ANSI S3.5-1997."

Speech Intelligibility Enhancement Using a Statistical Model of Clean Speech

Petko N. Petkov[1], *W. Bastiaan Kleijn*[1,2], *Gustav Eje Henter*[1]

[1]Sound and Image Processing Lab, School of Electrical Engineering,
KTH-Royal Institute of Technology, Stockholm, Sweden
[2]School of Engineering and Computer Science, Victoria University of Wellington,
Wellington, New Zealand

Abstract

Subjective speech intelligibility deteriorates when the speech is presented in a noisy environment. A trivial solution to the problem is to increase the volume of the signal. In practice, however, this approach leads to fatigue and dissatisfaction on the side of the listener. It may also lead to excessive levels of output power that cause non-linear distortions in the audio equipment. Alternatively, speech intelligibility can be enhanced by modifying the speech under an energy-preservation constraint.

An effective and powerful paradigm is to select the modification by optimizing the output of an objective intelligibility measure. We consider the application of this paradigm to an intelligibility measure related to the classification error probability in an automatic speech recognition system (ASR). We target, therefore, primarily the application to recorded and synthetic speech and assume that a transcription of the message is available when modifying the speech signal. The approach is general and can be applied to a broad range of modification strategies. The high operating level of the proposed measure suggests that modification selection is less influenced by mismatches between subjective and objective intelligibility.

We consider two modification strategies (arguably orthogonal to each other) in combination with the proposed objective measure. The choice of these modifications is motivated by findings resulting from the analysis of human behavior [1] and prior work on the effect of cue enhancement on improving subjective intelligibility [2]. In particular, we consider i) long-term spectral modifications and ii) vowel to consonant energy ratio adjustment at the word level. Both strategies have been applied in recent speech pre-emphasis algorithms using lower-level intelligibility measures [3, 4].

The objective measure we adopt is of the form:

$$\boldsymbol{c}^\star = \underset{\boldsymbol{c}}{\operatorname{argmax}} \sum_{j=1}^{J} w_j \log\left(\mathrm{p}\left(\boldsymbol{f}_j \mid \boldsymbol{m}_j, \boldsymbol{c}\right)\right)$$
$$\text{s.t. } \boldsymbol{c}^{\mathrm{T}}\bar{\boldsymbol{e}} = 1, \quad \boldsymbol{c} \geq 0, \tag{1}$$

where $\boldsymbol{f}_j,\ j \in \{1, \cdots, J\}$ are feature vectors extracted on a per-frame basis, $\boldsymbol{m}_j,\ j \in \{1, \cdots, J\}$ are Gaussian mixture models characterizing the features for particular phonetic units from the speech model in an ASR system pre-trained on clean speech, $\boldsymbol{c}$ represents the set of modification parameters, which can be viewed as gain factors in the spectral or the temporal domain depending on the modification, and w_j are weight factors that can be used, e.g., to manipulate the importance of the contribution of different phonetic groups. Energy preservation for the duration of the modification window is enforced by the equality constraint, which ensures that the sum of the normalized spectral-band or phone-unit energies $\bar{\boldsymbol{e}}^{\mathrm{T}} = [\bar{e}_1, \bar{e}_2, \cdots, \bar{e}_K]$ is preserved by the modification.

Subjective evaluation of the proposed approach was performed for an additive noise scenario (multi-speaker babble noise at $-3\mathrm{dB}$, SNR) using recorded speech and the spectral modification strategy at the word level. The results indicated significant and consistent improvement in subjective intelligibility for the modified over the original speech. Preliminary experiments with the temporal modification strategy revealed that it is possible to induce the desired behavior (relocating energy from vowels to consonants) and improve the intelligibility of the speech signal, as judged in informal subjective tests. Formal subjective evaluation of the temporal and the combination of the two strategies are planned to be performed shortly.

One specific challenge that needs to be overcome to ensure the robust performance of the proposed approach is related to the alignment between acoustic models and signal frames. While this information is *a priori* available for synthetic speech, it needs to be derived for recorded speech. We used forced alignment [5] between the transcription of an utterance and the clean speech waveform to achieve that. We observed that forced alignment often fails to locate correctly the individual phones and creates an error, which then propagates through the optimization process. A limitation of the approach was also established in regards to the spectral modification at the word level in relation to modifying fast speech. Edge effects together with sharp differences in the spectral gains between neighboring modification windows can effectively decrease the intelligibility of the modified below that of the original speech.

1. References

[1] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of Noise on Speech Production: Acoustic and Perceptual Analyses." *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, Sep 1988.

[2] V. Hazan and A. Simpson, "The Effect of Cue-Enhancement on Consonant Intelligibility in Noise: Speaker and Listener Effects," *Language and Speech*, vol. 43, pp. 273–294, 2000.

[3] B. Sauert and P. Vary, "Near End Listening Enhancement Optimized with Respect to Speech Intelligibility Index and Audio Power Limitations," in *Proc. Europ. Sig. Proc. Conf.*, 2010, pp. 1919–1923.

[4] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A Speech Preprocessing Strategy for Intelligibility Improvement in Noise Based on a Perceptual Distortion Measure," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2012, pp. 4061–4064.

[5] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2009.

High quality synthetic speech on a wide vocal effort continuum: Statistical parametric speech synthesis with glottal pulse library

Tuomo Raitio[1], *Antti Suni*[2], *Martti Vainio*[2], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[2]Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland
tuomo.raitio@aalto.fi, antti.suni@helsinki.fi

1. Introduction

Humans adapt their speech according to auditory environment in order to get the message delivered but without using unnecessary effort. Depending on the context, natural speech might vary from whisper to shouting. This vocal effort continuum is an integral part of human communication, but it is typically not utilized in machine-to-human communication. In order to produce contextually appropriate synthetic speech, the auditory environment and context must be taken into account and speech produced at a corresponding point in the vocal effort continuum.

2. Problem formulation

Modeling speech over a wide vocal effort continuum is not easy. In unit selection synthesis, this would require recording of various large databases along the continuum. In statistical parametric synthesis, two or more smaller databases recorded along the continuum can be used to adapt the normal voice. However, the quality of the adapted voices is not always adequate due to insufficient vocoder techniques, statistical averaging and small amount of data [1].

The problem with any speech synthesis system is that there are too little data, resulting in unseen contexts. However, if some contexts are separated into its components, all combination of the components need not to exist in the speech data. This property is utilized in the recently introduced hybrid unit selection/HMM-based system [2].

3. Hybrid unit selection/HMM-based system

A novel hybrid unit selection/HMM-based method [2], called Glottal Pulse Library technique, is based on using glottal inverse filtering for separating speech signal into a glottal source signal and a vocal tract filter. The estimated glottal source signal is segmented to individual glottal source pulses and parameterised into voice source features. Thus, in the synthesis stage, the excitation signal can be reconstructed by selecting the best matching pulses from the library according to the parameters generated by the HMM. The benefit of such a hybrid unit selection/HMM-based system is that the number of units required for natural sounding synthetic speech is very low since the two components, the glottal source and the vocal tract filter, are separated. Thus, only the varying context or modes of the voice source need to be stored into a pulse library, and the variation due to vocal tract filter is modeled by the HMM.

4. Results

Previously, we have shown that the glottal inverse filtering based vocoder [3] can successfully produce natural and very intelligible Lombard speech [4]. In this paper, we show that the glottal pulse library technique can successfully create a continuum from low to high vocal effort by using small pulse libraries along the continuum. We also explore new parameters for describing the glottal source pulses; we use traditional voice quality parameters, such as H1-H2 [5] and NAQ [6], as a target cost in the selection of the library pulses. Moreover, the pulse library method is faster than our previous implementation of the glottal inverse filtering based vocoder [3].

5. Acknowledgements

This research is supported by EU FP7 Simple4All, the Academy of Finland (projects 1128204, 1218259, 121252, 135003, LASTU), and MIDE UI-ART.

6. References

[1] Zen, H., Tokuda, K. and Black, A. W., “Statistical parametric speech synthesis”, Speech Commun., 51(11):1039–1064, 2009.

[2] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., “Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis”, ICASSP, 2011, pp. 4564–4567.

[3] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., “HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering”, IEEE Trans. on Audio, Speech, and Lang. Proc., 19(1):153–165, 2011.

[4] Raitio, T., Suni, A., Vainio, M. and Alku, p., “Analysis of HMM-based Lombard speech synthesis”, Interspeech, 2011, pp. 2781–2784.

[5] Titze, I. and Sundberg, J., “Vocal intensity in speakers and singers”, J. of the Acoustical Society of America 91(5):2936–2946, 1992.

[6] Alku, P., Bäckström, T. and Vilkman, E., “Normalized amplitude quotient for parametrization of the glottal flow”, J. of the Acoustical Society of America, 112(2):701–710, 2002.

Effect of expanding vs. reducing vowel contrast on adaptation to altered auditory feedback

Amélie Rochet-Capellan[1], David J. Ostry[2]

[1] ZAS, Berlin, Germany, [2] McGill University, Montreal, Canada
ameliecapellan@free.fr, david.ostry@mcgill.ca

Abstract

Speakers modify their speech in order to compensate for a perturbation in their auditory feedback. Auditory-motor adaptations can be studied by applying specific real time perturbations to a speaker's auditory feedback. For example, when a speaker produces vowels, it is possible to change the value of the first formant (F1) in the acoustic signal and to play it back in real time to the speaker through headphones. Speakers learn to compensate for this perturbation by changing F1 in their production in a direction opposite to the perturbation [1, 2, 3]. This compensation lasts for a period of time after the perturbation is removed, which shows that motor learning occurred.

Most of the studies on the adaptation of F1 perturbations have used a single perturbation consisting of either increasing or decreasing F1 in speakers' auditory feedback. Recently [4], using a similar protocol, we showed that speakers can also simultaneously learn several auditory-motor transformations when each perturbation is applied to the production of a vowel in a given word. Hence, when speakers are trained to produce the words "head" and "had" repetitively and in random order, receiving an upwards perturbation of F1 for "head" and a downwards perturbation for "had", they progressively decrease the F1 value in their "head" productions while at the same time, they increase the F1 value in "had" . We interpreted these results as evidence that motor learning in speech is linked to a given word or that sensori-motor learning is quite specific.

The method we developed (applying different perturbations to different vowels in a same training session) is also a way to modify the contrast between two utterances in subjects' auditory feedback. For example, when F1 is shifted upwards in "head" and downwards in "had", this makes the auditory feedback of "head" and "had" more similar. In contrast, the reverse perturbations make the two utterances more distinct. Hence, in our previous work, the adaptations observed could also be interpreted as an effort to preserve "head"-"had" contrast.

In the current study, we re-analyzed the adaptations that we observed in our previous work as a change in the F1-F2 distance between the two vowels, rather than an adaptation of F1 for each vowel. We also compared these adaptations to those of two other groups of speakers trained in similar conditions but received perturbations only for "head" utterances, not for "had" (which is another way to change "head"-"had" contrast).

We observed that, as in our previous results, subjects learnt to change the motor control for each utterance in different ways. However, analysing the adaptation in term of relations between the two words shows that speakers also adapt to maintain the contrast between the two utterances. Thus, speech-specifics constraints, such as contrast between two target sounds, also play a role in auditory-motor adaptation.

References

[1] Houde JF and Jordan MI (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. Journal of Speech and Hearing Language. 45, 295 – 310.

[2] Purcell, D.W. and Munhall, K.G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. Journal of the Acoustical Society of America, 120, 966-977.

[3] Villacorta VM, Perkell JS et Guenther FH (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. Journal of the Acoustical Society of America;122(4):2306-19.

[4] Rochet-Capellan A, Ostry DJ (2011). Simultaneous acquisition of multiple sensorimotor transformations in speech. Journal of Neuroscience, 31:2648-2655.

Does listeners' breathing change according to speaker and to loudness?

Amélie Rochet-Capellan[1], Susanne Fuchs[1], Leonardo Lancia[2], Pascal Perrier[3]
[1] ZAS, Berlin, Germany, [2] MPI, Leipzig, Germany, [3] GIPSA-lab/DPC Grenoble, France
ameliecapellan@free.fr, fuchs@zas.gwz-berlin.de, leonardo_lancia@eva.mpg.de,
Pascal.Perrier@gipsa-lab.grenoble-inp.fr

Abstract

Breathing has specific kinematics during speech production that changes, for example, according to loudness [1, 2] or to the speakers' gender [3]. Breathing also shows a specific profile during listening to speech, that is intermediate between the profile of quiet (or rest) breathing and the profile of breathing during speech production [2, 4]. In dialogue speaker' and listener' breathing synchronize at the time of turn-taking [4]. This suggests a mutual adaptation of speaker' and listener' breathing, which could be a part of the interactive speaker-listener alignment process that occurs in communication [5]. In this framework we evaluated if breathing changes during listening to speech depend on the speaker who produces the speech and/or on the loudness of its speech signal.

We recorded acoustic and breathing from two native speakers of German (a male and a female) while they were reading short texts with either a normal or a loud (~+10dB) acoustic level. The two readers had different breathing kinematics. The duration of the breathing cycle (inhalation+exhalation) was smaller for the female than for the male. The amplitude of the breathing cycle increased with loudness for both subjects while the frequency decreased for the male speaker but remained unchanged for the female. Then, we monitored breathing for 26 native females speakers of German while they were listening to the readers' recordings audio played back via loudspeakers. They listened either to the male or to the female speaker, starting either with the normal (5 texts) or with the loud readings (5 texts). After listening to each text, listeners had to briefly summarize it. We analyzed breathing kinematics during listening according to the reader (male vs. female), to the loudness condition (normal vs. loud) and to the order (normal first vs. loud first).

As in previous studies, we found that breathing kinematics during speech was different than during quiet breathing [2]. This shows that breathing was involved or at least affected by the listening process. The comparison between the different listening conditions showed that the duration and the amplitude of the breathing cycles were greater when listening to normal speech than when listening to loud speech. The changes in duration due to changes in loudness were similar for listeners to both readers. In contrast, an increase in amplitude was observed for listeners to the male reader while listeners to the female reader tend to show the reverse effect. Finally, we observed some effects of the presentation order of conditions, especially for the male reader.

Altogether, these preliminary results show that listeners' breathing is sensitive to the reader and to the loudness of the reader's speech. This sensitivity could be a physiological reaction, as breathing is closely linked with heartbeats and emotional state. However, the changes in listeners' breathing could also be due to an adaptation to specific characteristics of the reader's voice and/or rhythms. It could also result from a speaker-listener's breathing coupling, as it has been observed for body movements in dialogue or for brain activity during listening [6, 7, 8].

References

[1] Huber JE, Chandrasekaran B, Wolstencroft JJ (2005). Changes to respiratory mechanisms during speech as a result of different cues to increase loudness. J. Appl. Physiol., 98: 2177–2184.
[2] Conrad B, Schonle P (1979). Speech and respiration. Arch Psychiatr Nervenkr, 226: 251–268.
[3] Hoit JD, Hixon TJ, Altman ME, Morgan WJ (1989). Speech breathing in women. J Speech Hear Res, 32: 353–365.
[4] McFarland DH (2001). Respiratory markers of conversational interaction. J. Speech Lang. Hear. Res., 44: 128–143.
[5] Pickering MJ, Garrod S (2004). Toward a mechanistic psychology of dialogue. Behav Brain Sci, 27, 2:169-90; discussion 190-226.
[6] Shockley, K, Santana, MV, Fowler, CA (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. J Exp Psychol Hum Percept Perform, 29, 2:326-32.
[7] Schmidt, RC, Fitzpatrick, P, Caron, R, Mergeche, J (2011). Understanding social motor coordination. Hum Mov Sci, 30, 5:834-45.
[8] Stephens, G., Silbert, L., and Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. Proc Natl Acad Sci, 107(32):14425–14430.

Expanding the vowel space – direct vocal tract measurement with ultrasound tongue imaging

James Scobbie

Queen Margaret University, Edinburgh, UK

jscobbie@qmu.ac.uk

Abstract

In noisy conditions, speakers adapt their speech production in a number of ways, which can be usefully studied through acoustic analysis of the speech output or perceptual testing of this output. The noisy conditions can affect the speaker, the speaker and listener, or the listener alone, and in all cases, the speaker may use Lombard speech. This is a type of effortful speech causing a variety of effects such as an expanded vowel space. The hypothesis from the source-filter theory of the vocal tract is that speakers are altering their supralaryngeal articulations in order to enhance formant values, thereby making their vowels more perceptibly distinct. Our focus is on how speakers enhance their speech towards a listener whose hearing is artificially masked, as a model of how speakers may enhance their speech when talking to hearing-impaired listeners.

We will measure speech articulation directly, using Ultrasound Tongue Imaging and a lip-jaw camera synchronised with the acoustics, in order to describe the ways in which different speakers expand their vowel space. The camera is mounted on a headset worn by the speaker also it also holds the ultrasound probe steady. Ultrasound provides a mid-sagittal tongue surface image from near the tip of the tongue right down to the root, near the larynx, and is cheap, quiet, and provides dynamic images at a high frame rate. We will use Scottish English for this study because it has 9 monophthongal vowels which provide an excellent sampling of the vowel space.

Two conditions will elicit clear read speech vs. Lombard speech. Speakers will first read a word list to sample vowels in the neutral context (multiple repetitions of haw, hoe, who, him, etc). There will be no background noise. In such an experimental situation, speakers produce relatively clear but neutral speech. In the second condition, intended to elicit Lombard speech, a listener will come into the room and wear headphones with masking speech babble played at 40dB, and the word lists will be repeated by the speaker, in a different random order. The listener will be given the (genuine) task of transcribing each word accurately. They will be able to ask the speaker to repeat unclear tokens. This condition makes the speaker enhance their speech production, which the ultrasound-camera system will record, along with acoustics. Since only the listener experiences the masking noise, the audio recordings are also available for a comparative analysis with the first condition.

To preserve speaker comfort and because articulatory analysis is time-consuming, we will be limited to analysing 5 speakers, with approximately 6 tokens of each vowel in each condition.

We will analyse the difference between the two production conditions by tracing the tongue surface in each vowel and calculating the average position of the target, and comparing differences in vowel shape and location across conditions. In addition, the entire vowel articulation space can be measured, as the physiological area bounded by extreme low, high, front and back vowels. The physical area exploited for vowel production is hypothesised to be larger in Lombard speech.

We will also undertake analysis of the oral cross-sectional area of the lips, but the hypotheses on this topic are unclear. We would expect to find trade-offs between tongue and lip articulations to create formant effects in case the listener is able to see the speaker (as in this case), or not, and this could be explored in future research. Implications for theories of speech production and the interplay of audio-visual cues to speech perception will be discussed.

Intelligibility and Production in Greek Hearing-Impaired Speech

Anna Sfakianaki [1], *Katerina Nicolaidis* [1], *Areti Okalidou* [2]

[1] Department of Theoretical and Applied Linguistics, Aristotle University of Thessaloniki, Greece
[2] Department of Educational and Social Policy, University of Macedonia, Greece
asfakian@enl.auth.gr, knicol@enl.auth.gr, okalidou@uom.gr

Abstract

Talkers with prelingual profound hearing impairment (PHI) develop speech without adequate auditory feedback. Consistent with the DIVA model of speech motor planning, auditory feedback plays a key role in the development and continual tuning of the speech production mechanism [1], [2]. The purpose of the present study was to examine the speech intelligibility of Greek adults with PHI in relation to vowel system characteristics, such as duration, position in the acoustic space and token-to-token variability.

Index Terms: hearing-impaired talker, normal-hearing listener, speech intelligibility, vowel space, Greek

1. Introduction

Speech intelligibility refers to the degree to which listeners receive the talker's intended verbal message. Intelligibility levels of talkers with PHI are characterized by great variability [3], [4]. Intelligibility level is a useful indicator of oral communication abilities, but its relation to speech production characteristics needs to be explored as well, so as to design effective remediation for individuals with hearing loss [5]. This study investigates (a) selected acoustic properties of the vowel system of Greek adult talkers with PHI using conventional hearing aids and (b) the relationship between vocalic acoustic characteristics and PHI speech intelligibility level as judged by naïve listeners with normal hearing.

2. Method

Two experiments were carried out, one measuring speech intelligibility and the other looking into selected acoustic characteristics of vowels.

For the intelligibility experiment, 101 words and 25 sentences were recorded by five men and five women with prelingual profound hearing loss ranging from 91 to 105 dB HL, who had not received cochlear implants but made continuous use of hearing aids since the diagnosis (before the age of 4). The material was judged by 60 naïve listeners with normal hearing, and scored according to [6].

For the production experiment, symmetrical disyllables of the form /pVpV/ (V= [i, a, u]) with stress placed on the first or second syllable, embedded in a carrier phrase, were uttered by the nine talkers with intelligible PHI speech and by five talkers with normal hearing, two men and three women, serving as a control group. Formants F_1 and F_2 at vowel midpoint as well as vowel duration were measured in both syllable positions.

3. Results

The results of the intelligibility experiment showed that intelligibility level ranged from medium to very high, with the exception of one talker with unintelligible speech. Additionally, the higher identification rate of words in sentences vs. in isolation indicates that context is an important factor in successful communication, in agreement with [7]. The acoustic analysis revealed reduced vocalic contrast, higher acoustic variability and longer vowel durations than normal. The combined results of the two experiments suggest an inverse relationship between overall speech intelligibility and acoustic space mainly due to a more anterior production of [u]. However, acoustic variability and duration did not seem to correlate with intelligibility level.

4. Conclusions

Talkers with PHI are seriously disadvantaged listeners. Their speech production mechanism matures under severe restrictions due to limited ability to monitor their own speech as well as inadequate opportunity for adaptation to external conditions during verbal communication. The results of the present study provide evidence for the detrimental effect of reduced vocalic contrast on speech intelligibility and hold the premise that the exploration of perceptuo-production links underlying the communication of hearing-impaired talkers and normally-hearing listeners is necessary in order to develop improved technological and remedial strategies for successful communication.

5. References

[1] Guenther, F. H., "A neural network model of speech acquisition and motor equivalent speech production", Biological Cybernetics, 72:43-53, 1994.

[2] Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., Wilhelms-Tricarico, R. and Zandipour, M., "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss", Journal of Phonetics, 28:233-272, 2000.

[3] Smith, C. R., "Residual hearing and speech production in deaf children", Journal of Speech and Hearing Research, 18:773-794, 1975.

[4] Tobey, E., Geers, A., Douek, B., Perrin, J., Skellett, R., Brenner, C. and Toretta, G., "Factors associated with speech intelligibility in children with cochlear implants", Annals of Otology, Rhinology, and Laryngology, Suppl 185, 109(12):28-30, 2000.

[5] Metz, D. E., Samar, V. J., Schiavetti, N., Sitler, R. and Whitehead, R. L., "Acoustic dimensions of hearing impaired speakers' intelligibility", Journal of Speech and Hearing Research, 17:386-398, 1985.

[6] Osberger, M. J., Maso, M. and Sam, L. K., "Speech intelligibility of children with cochlear implants, tactile aids, or hearing aids", Journal of Speech and Hearing Research, 36:186-203, 1993.

[7] McGarr, N., "The effect of context on the intelligibility of hearing and deaf children's speech", Language and Speech, 24(3):255-264, 1981.

WinkTalk: a multimodal speech synthesis interface linking facial expressions to expressive synthetic voices

Éva Székely, Zeeshan Ahmed, João P. Cabral, Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics, University College Dublin
Dublin, Ireland
{eva.szekely|zeeshan.ahmed}@ucdconnect.ie,{joao.cabral|julie.berndsen}@ucd.ie

1. Introduction

During a human verbal communication process, expressive features of face and speech are congruent, operating in a synchronised manner [1]. Facial expressions and expressive speech styles often help conveying the emotional message of the speaker that is only partially contained in the words. The application described in this abstract is aimed to make use of this synchrony by applying facial expressions as a control over the expressive features of synthetic speech in a situation where a person uses a laptop equipped with a web camera and a speech synthesiser to communicate with another person in the room. The system is currently a research prototype in progress. The goal of the WinkTalk system is to respond to the need of integrated multimodality in interactions of users of augmentative and alternative communication (AAC) applications [2], on a proof of concept level. Being able to correctly link facial expression to synthetic voice style is a step forward for providing a more intuitive way of controlling the expressiveness of the synthetic speech. The application uses a web camera and face detection software [3] to analyse the users facial expression while sending a message to the synthesiser. In an attempt to convey the intended non-linguistic content of the message, the system selects an expressive synthetic voice that matches the type and intensity of the detected facial expression of the user.

2. WinkTalk system components

2.1. Facial expression analysis

The facial expression analysis in WinkTalk is conducted by SHORE, a real time face detection engine freely available for academic purposes of demonstration and evaluation [3]. To detect faces and expressions, SHORE analyses local structure features in (series of) images that are computed with a modified census transform [4]. The face detection outputs scores for four distinct facial expressions, happy, sad, angry and surprised, with an indication of the intensity of the expression.

2.2. Synthetic voices

The synthesiser component of the application uses three expressive HMM-based synthetic voices [5] of a middle aged American male. The voices have been built from three different sections of an audiobook corpus, each featuring a different expressive voice style. Perceptual experiments have shown that the three voices can be characterised on an expressiveness gradient: from calm (A voice), through intense (B voice) to very intense (C voice) expressions [6].

2.3. Mapping of face and voice

The application uses a mapping of the synthetic voices to the type and intensity of different facial expressions as analysed by the face detection software. The mapping rules have been defined by a combination of theoretical considerations based on arousal levels of underlying basic emotions for each facial expression and empirical data collected in user evaluations conducted on a balanced dataset of visual and auditory stimuli. Beyond that, the application contains a personalisation component, where the mapping rules between facial expression and speech are adjusted to take into account personal differences between users.

3. Evaluation and Conclusions

The WinkTalk system has been evaluated with an interactive evaluation session involving 10 subjects acting out pre-scripted dialogs. The evaluation has shown that there is a general preference to manual selection of expressive voices, 90% of the participants described facial expression control as a valuable addition to the simulated augmented communication process.

4. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin (UCD). The authors would also like to thank Shannon Hennig (IIT), Nick Campbell and the Speech Communication Lab (TCD) for their invaluable help with the interactive evaluation.

5. References

[1] Campbell, N. “Multimodal processing of discourse information; the effect of synchrony,” Proceedings of International Symposium on Universal Communication, vol. 0, pp. 1215. 2008.

[2] Higginbotham, D. J. “Humanizing Vox Artificialis: The Role of Speech Synthesis in Augmentative and Alternative Communication,” in Computer Synthesized Speech Technologies: Tools for Aiding Impairment, J. Mullennix and S. Stern, Eds. IGI Global, 2010, pp. 50-70.

[3] http://www.iis.fraunhofer.de/en/bf/bsy/fue/isyst retr. 27.01.2012

[4] Kueblbeck., C. and Ernst, A. “Face detection and tracking in video sequences using the modified census transformation”, Journal on Image and Vision Computing, vol. 24, issue 6, 2006.

[5] HTS-2.1 toolkit “HMM-based speech synthesis system version 2.1”, http://hts.sp.nitech.ac.jp, 2008.

[6] Hennig, S., Székely, É., Carson-Berndsen, J., Chellali, R. (accepted) “Listener evaluation of an expressiveness scale in speech synthesis for conversational phrases: implications for AAC.” Proceedings of ISAAC, 2012.

Optimal frequency filtering for speech intelligibility boosting under a constant energy constraint

Yan Tang[1], *Martin Cooke*[2,1], *Petko N. Petkov*[3]

[1]Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain
[2]Ikerbasque (Basque Science Foundation), Bilbao, Spain
[3]Sound and Image Processing laboratory, School of Electrical Engineering, KTH-Royal Institute of Technology, Stockholm, Sweden

y.tang@laslab.org, m.cooke@ikerbasque.org, petkov@kth.se

Natural or synthetic speech is increasingly used in less-than-ideal listening conditions. While talkers have a range of potential strategies available to promote successful communication by adapting to the listener's context, speech output technology has in the past been largely insensitive to the changing needs of the listener. Recently, context-sensitive speech output algorithms have been proposed and evaluated [1, 2, 3, 4, 5] for both recorded and synthetic speech. These approaches typically modify the speech signal to maintain speech intelligibility without increasing signal level or duration. Two problems need to be solved to deliver benefits in context-sensitive speech output technology. First, speech modifications must be found which lead to intelligibility increases. Second, the acoustic context (e.g. background noise type, level) must be known, predictable or estimated with sufficient accuracy to enable speech modification algorithms to make optimal adjustments. In [6] we evaluated a range of modification techniques which reallocated speech energy across time and frequency while preserving overall signal-to-noise ratio (SNR), demonstrating substantial listener benefits. However, the most successful modifications required detailed local noise estimates over time which may be difficult to deliver in practice. An intermediate approach to the use of context is to estimate noise descriptors, and to use these to select a modification which has been optimised offline.

The current study focuses on the problem of offline optimisation of speech modifications designed to promote intelligibility in the context of different noise types at a range of SNRs. In this initial study, modifications are restricted to stationary spectral reweightings under globally-constant energy and duration constraints. Frequency band weights were selected using a genetic algorithm-based optimisation procedure [7], with glimpse proportion [8] as an objective intelligibility metric, for a range of noise types (competing talker, speech-shaped, speech-modulated, high-pass, low-pass and white noise) and noise levels producing global signal-to-noise ratios in the range +10 to -10 dB. Speech and noise signals were filtered into 58 frequency bands whose bandwidths were based on auditory filters in the range 50-8000 Hz, to later give a high quality reconstructed signal. One unanticipated outcome was the consistent discovery of sparse, highly-selective spectral energy weightings, particularly in high noise conditions (e.g. figure 1). With speech modified by applying optimal spectral weightings in a subjective test, listeners were able to identify significantly more words in sentences in the presence of stationary noise and competing speech maskers, with increases of up to 15 percentage points. These findings suggest that context-dependent speech output can be used to maintain intelligibility at lower sound output levels.

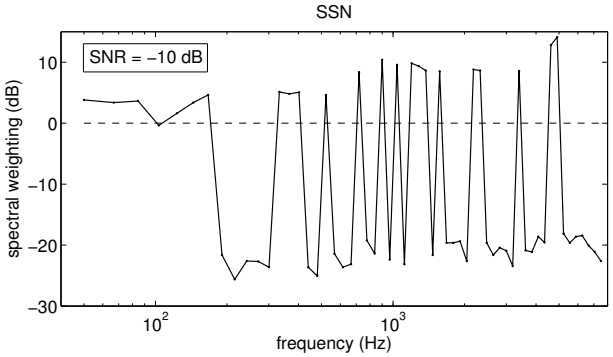


Figure 1: *GA-optimised weighting for speech-shaped noise at SNR = -10 dB.*

Acknowledgement: This study was supported by the LISTA Project, funded by the EU Future and Emerging Technologies programme (grant number 256230).

[1] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 1636–1639.

[2] S. D. Yoo, J. R. Boston, A. El-Jaroudi, C. Li, J. D. Durrant, K. Kovacyk, and S. S., "Speech signal modification to increase intelligibility in noisy environments," *J. Acoust. Soc. Am.*, vol. 122, no. 2, pp. 1138–1149, 2007.

[3] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *Proc. EUSIPCO-2010*, Aalborg, Denmark, 2010, pp. 1919–1923.

[4] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 1636–1639.

[5] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proc. Interspeech*, Florence, Italy, 2011.

[6] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 345–348.

[7] J. H. Holland, *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975, vol. Ann Arbor, no. 53.

[8] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

Effect of level and type of noise on focus related prosody

Martti Vainio[1], *Antti Suni*[1], *Anja Arnhold*[1,3], *Tuomo Raitio*[2], *Henri Seijo*[2], *Juhani Järvikivi*[4],Daniel Aalto[1,2], and *Paavo Alku*[2]

[1]Institute of Behavioural Sciences (SigMe Group), University of Helsinki, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Finland
[3]Department of Cognitive Linguistics, Goethe-University Frankfurt am Main, Germany
[4]Language Acquisition and Language Processing Lab, NTNU, Norway
martti.vainio@helsinki.fi

Speakers automatically raise their voice when forced to speak in environmental noise or when the normal feedback mechanism is disturbed. Raising one's voice consists of various physiological means that have different consequences on the phonetic realisation of speech. Typically the speakers' f_0 is higher – probably as a consequence of added sub-glottal pressure – and the mode of vocal fold vibration is more pressed in order to decrease the slope of the glottal voice-source spectrum. The adaptation of speech to noise in order to increase the signal-to-noise ratio is called the Lombard effect or Lombard reflex to illustrate its involuntary nature [1]. However, the knowledge regarding Lombard speech is fairly general in nature and not very much is known about how the reflex influences prosodic changes that are due to specific communicative needs such as signalling prosodic focus.

We recorded 21 speakers producing utterances with different information focus conditions in three types of noise with four noise levels. The purpose of the study was to see whether speakers vary their means of producing prosodic focus as a function of both noise level and type. The analysed utterances were replies to three types of questions designed to elicit either a broad focus or two types of narrow focus in simple three word utterances. The typical prosodic patterns for the three focus conditions are well-known for Finnish, which allows us to compare Lombard speech to normal speech in a controlled manner. The three types of noise were: white noise, babble noise, and a 1 kHz low-pass noise. The noises were scaled for equal loudness on three different levels corresponding to approximately 60, 70 and 80 dB(A) sound pressure levels.

We analyzed the produced utterances with regard to f_0, *duration*, *voice source features*, *formants* and *intensity*. Only pitch related features are presented here. The pitch contours were analyzed in terms of three different points per word: the pitch maximum (peak) and the minima left and right of it (valleys). Thus, there are nine potential values for each utterance. The contours clearly follow the typical shapes associated with different focus conditions in Finnish [2, 3]; i.e, the narrowly focused word has a higher peak and post-focal words have lower peaks but are not altogether deaccented. The verbs also have a rising-falling shape, but with a markedly lower magnitude [4]. For further analyses the f_0 values were transformed to semitones (re 100 Hz).

The f_0 *expansion* was calculated from the nine f_0 values per sentence (also in semitones) by adding the absolute differences. This can be expressed in terms of an integral based on the Bounded Variation (BV) norm:

$$\text{Expansion}(f_0) = \int_{T_{beg}}^{T_{end}} \left| \frac{df_0(s)}{dt} \right| ds \qquad (1)$$

where $f_0(t)$ is the fundamental frequency at a given time point and T_{beg} and T_{end} are the beginning and end times of the utterance.

Statistical analyses with linear mixed-effects models show that with regard to mean f_0 the noise levels differ significantly. The low-pass noise has a significantly lower mean f_0. As can be expected from the results of previous studies, the different focus types are also different from each other: i.e, the f_0 is generally lower when the narrow focus occurs on the first word and higher when it occurs on the last word. There is also a significant low-pass-noise:noiselevel3 interaction showing that the f_0 level is increased less in high level low-pass noise. With regard to f_0 expansion the results show that the contours are significantly influenced by the focus type as well as noise levels. The low-pass noise, however, does not differ from babble noise, but the contours are again more expanded in white noise (t = 3.64). This is also shown in the white-noise:level interactions. There are no significant gender differences.

In summary, the analysis of the data shows that, regardless of the increase in f_0, the typical intonation contours are still produced. Also, and as can be expected, the f_0 is raised as a function of noise level. Moreover, the intonation contour is expanded as the f_0 gets higher; the louder the noise, the more expanded the contour is. In addition the noise type affects the contour in different ways and and there are interesting level-type interactions. The typical utterance-final creaky voice is also not as prevalent in Lombard speech as it is normally and disappears altogether in severely noisy conditions.

1. References

[1] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 101-119, p. 25, 1911.

[2] M. Vainio and J. Järvikivi, "Focus in production: Tonal shape, intensity and word order," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. EL55–EL61, 2007. [Online]. Available: http://link.aip.org/link/?JAS/121/EL55/1

[3] A. Arnhold, "Multiple prosodic parameters signaling information structure: Parallel focus marking in finnish," in *Proceedings of the International Congress of Phonetic Sciences (ICPhS), Hong Kong*, 2011.

[4] A. Arnhold, M. Vainio, A. Suni, and J. Järvikivi, "Intonation of Finnish verbs," in *Speech Prosody 2010-Fifth International Conference*, 2010.

Using an intelligibility measure to create noise robust cepstral coefficients for HMM-based speech synthesis

Cassia Valentini-Botinhao[1], *Yan Tang*[2], *Junichi Yamagishi*[1], *Simon King*[1]

[1] The Centre for Speech Technology Research, University of Edinburgh, UK
[2]Language and Speech Laboratory, Universidad del País Vasco, Spain

C.Valentini-Botinhao@sms.ed.ac.uk, y.tang@laslab.org,
jyamagis@inf.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

The aim of this work is to increase intelligibility of HMM-based synthetic speech in noisy environments by modifying clean synthetic speech given that noise is known. For that purpose we need a measure for intelligibility of speech in noise that can automatically define the sort of modifications that we need to apply. In previous experiments [1] we have observed that spectrum envelope modifications can have a significant positive impact on the intelligibility of HMM-generated synthetic speech in noise and that the Glimpse proportion measure (GP) [2] is highly correlated with subjective scores under those circumstances.

We have then introduced a method for cepstral coefficient extraction that modifies spectrum envelope based on the GP measure. The GP accounts only for the effect of additive noise, not requiring a reference unmodified speech signal to produce a intelligibility prediction. To control the amount of distortions introduced by the modification we extract cepstral coefficients using an optimization criterion with two terms. The first term accounts for the minimization of the mismatch between natural speech periodogram and magnitude spectrum as modeled by cepstral coefficient, the current criterion used for cepstral coefficient extraction performed at the training stage of the HMM-based speech synthesis framework [3]. The second term accounts for the maximization of an approximated analytical and differentiable version of the GP measure. Using this method we found significant intelligibility gains however not for all tested noise types which indicates that we need a more effective method for controlling distortions [4].

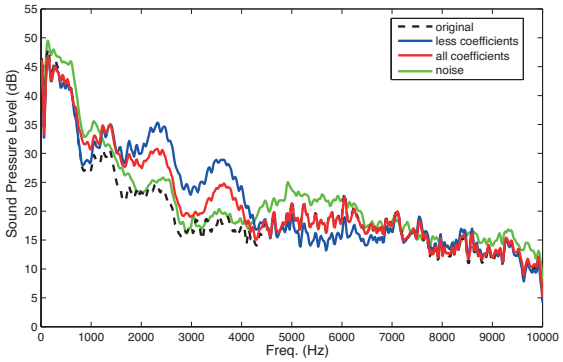


Figure 1: *Long term average spectrum of the original, modified less coefficients (8 coefficients) and modified all coefficients (59) for speech-shaped noise.*

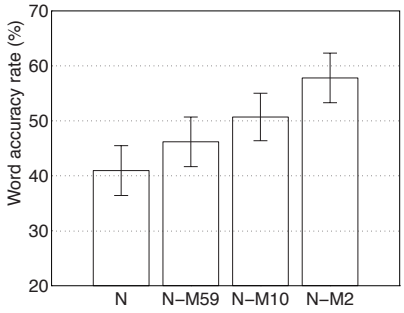


Figure 2: *Word accuracy scores: original (N) and proposed when all (N-M59), 10 (N-M10) and 2 (N-M2) coefficients were modified.*

In this work we propose to limit the frequency resolution of the modifications, and therefore the amount of distortions, by altering only the first few cepstral coefficients, known to be responsible for the coarse frequency resolution of the spectrum. Fig.1 shows the long term average spectrum of original and modified speech, where we can see the effect that limiting the degrees for freedom has on the spectrum envelope. Listening experiments results as shown in Fig.2 indicates that when we modify less coefficients we can improve intelligibility even further.

1. References

[1] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proc. Interspeech*, Florence, Italy, August 2011.

[2] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[3] K. Tokuda, T. Kobayashi, and S. Imai, "Adaptive cepstral analysis of speech," *IEEE Trans. Speech and Audio Processing*, vol. SA-3, no. 6, pp. 481–489, Nov. 1995.

[4] C. Valentini-Botinhao, R. Maia, J. Yamagishi, S. King, and H. Zen, "Cepstral analysis based on the Glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise," in *Proc. ICASSP*, Kyoto, Japan, March 2012.

The role of durational changes in the Lombard speech advantage

Julián Villegas[1,2], *Martin Cooke*[1,2], *and Catherine Mayo*[3]

[1]Ikerbasque (Basque Science Foundation), Spain
[2]Language and Speech Laboratory, Universidad del Pais Vasco, Spain
[3]Centre for Speech Technology Research, University of Edinburgh, UK
j.villegas@laslab.org,m.cooke@ikerbasque.org,catherin@inf.ed.ac.uk

Abstract

Speech produced in the presence of noise—Lombard speech (LS)—has been found to be more intelligible than 'normal' speech when presented in equivalent amounts of noise [1, 2]. However, the origin of the LS advantage remains unclear. Part of the benefit appears to stem from spectral changes in LS which shift energy into the 0.6–3 kHz region (See Fig. 1) where it better escapes energetic masking by speech-shaped noise. Other parameters which show changes in LS include f_0 and duration. Lu & Cooke [3] modified the mean f_0 and spectrum (both independently and jointly) of normal speech, demonstrating a clear advantage of spectral modification but no effect of f_0.

The current study extended [3] in two directions. First, durational modifications to reflect differences between normal and LS were included. Second, as well as global (per utterance) changes, local (per frame) modifications were applied. Four male and four female talkers produced simple sentences containing spoken letter and number keywords in quiet and in intense speech-shaped noise (96 dB SPL). A perception experiment explored global vs. local modifications of spectral and durational parameters applied independently.

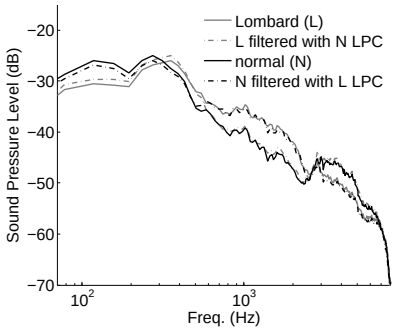


Figure 1: LTAS of the combined sentences used in the global spectral modifications.

Taken across all talkers, durational changes to normal speech produced no intelligibility benefit, whether applied globally (i.e., linear stretching or compression) or locally (using dynamic time warping to align normal and Lombard frames), spectral changes based on global or local modification were equally beneficial (See Fig. 2). However, when sentences were partitioned in two groups according to their speech rate in noise (See Fig. 3), some effect of durational modifications was observed: normal speech modified globally to the faster speech rate group was significantly less intelligible than unmodified speech, while conversely for the group with slower LS a small intelligibility benefit was present.

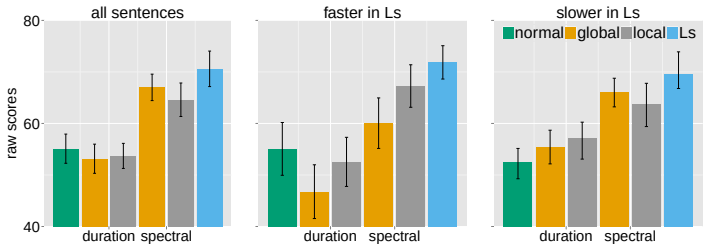
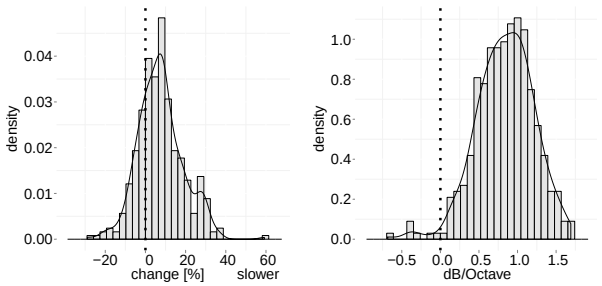


Figure 2: Raw intelligibility scores. Error bars correspond to the 95 % confidence interval.

These findings suggest that durational differences between normal and LS can affect intelligibility, but the benefit or otherwise depends on individual differences in speech rate.



(a) duration ($M = 7.58$) (b) spectral tilt ($M = 0.84$)

Figure 3: Differences between LS relative to normal.

Acknowledgements. The authors thank EU Future and Emerging Technology (FET–OPEN) Project LISTA (The Listening Talker) and staff at the Centre for Speech Technology Research of the University of Edinburgh for their collaboration running the experiment.

[1] J. J. Dreher and J. O'Neill, "Effects of ambient noise on speaker intelligibility for words and phrases," *J. Acoust. Soc. Am.*, vol. 29, no. 12, pp. 1320–1323, 1957.

[2] W. V. Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[3] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Comm.*, vol. 51, pp. 1253–1262, 2009.

Improving speech intelligibility in noise environments by spectral shaping and dynamic range compression

Tudor-Catalin Zorila[1], *Yannis Stylianou*[1,2]

[1]Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Greece
[2]Computer Science Department, University of Crete, Heraklion, Crete, Greece

tudorcatalin.zorila@yahoo.com, yannis@csd.uoc.gr

1. Introduction

Speech produced under real conditions (not a recording studio, nor a quiet room) is not always equally intelligible due to the presence of background noise. This noise may mask part of the speech signal such that not all speech information is available to the listener. The ability to detect speech in noise plays a significant role in our communication with others. In this work we suggest the use of a non-parametric way to improve the intelligibility of speech under adverse noisy conditions by modifying the speech signal accordingly.

2. Method

The suggested system contains two subsystems: (i) Spectral Shaping (SS) and (ii) Dynamic Range Compressor.

The goal of Spectral Shaping is to increase the "crisp" and "clean" quality of the speech signal, and therefore improve the intelligibility of speech even in clear (not-noisy) conditions. For this, both adaptive and fixed spectral shaping operators are used. The adaptive spectral shaping takes into account the probability of voicing given a speech frame, while the fixed spectral shaping is independent of the probability of voicing. The adaptive spectral shaping consists of (i) adaptive sharpening where the formant information is enhanced, and (ii) an adaptive pre-emphasis filter. The adaptive (to the probability of voicing) characteristic of the suggested system is important for not introducing artifacts in the processed signal especially in fricatives, silence or other "quiet" areas of speech. The purpose of the fixed (non-adaptive) spectral shaping is to protect the speech signal from low-pass operations during the reproduction of the signal.

The output of the Spectral Shaping system is the input to the Dynamic Range Compressor (DRC). DRC has a dynamic and a static stage. During the dynamic stage, the envelope of the signal is dynamically compressed with 2ms release time constant and almost instantaneous attack time constant. The signal envelope is based on the Hilbert transform and a moving average operator with order determined by the average pitch of speaker's gender. After the dynamic compression of the signal envelope, a static amplitude compression is applied. During the static amplitude compression, the 0 dB reference level is a key element in forming the Input/Output Envelope Characteristics (IOEC). For the current system this was set to 0.3 the peak of the signal.

The whole system is based on a frame-by-frame analysis and synthesis. In each frame the magnitude spectrum is computed using FFT and then manipulated in the way mentioned above. Overlap and add is then used to reconstruct the modified signal. The whole process is very fast and can run in real time.

3. Results

For testing the system, we used the first 20 Harvard sentences and two types of noise: Speech Shaped Noise (SSN) at SNR: -9dB, -4 dB and 1 dB, and Competing Speaker noise (CS) at SNR: -21 dB, -14 dB, -7 dB. For evaluation and comparison purposes, the extended SII suggested in [1] and the frequency dependent SNR recovery system suggested in [2], were implemented. Fig. 1 shows the results in terms of SII for the original signal without modifications (Orig), the suggested system (SS-DRC) and the system presented in [2] (referred to as SNR-R). Performance of sub-systems (Spectral Shaping (SS), Dynamic Range Compression (DRC)) is also shown. The final system combines SS and DRC in a cascade form. Overall, the suggested system (SS-DRC) outperforms SNR-R for all SNR levels and for both types of noise. All modified signals (either modified by SNR-R or SS-DRC) report better SII score than the non-modified signals.

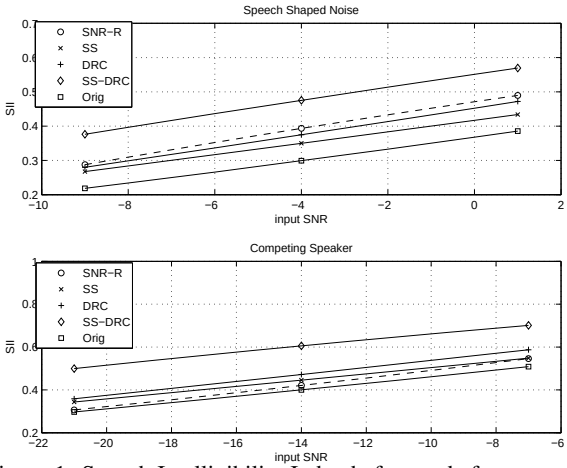


Figure 1: Speech Intelligibility Index before and after processing.

4. References

[1] K. Rhebergen and N. Versfeld, "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *JASA*, vol. 117, no. 4, pp. 2181–2192, 2005.

[2] B. Sauert and P. Vary, "Near end listeniing enhancement: Speech intelligibility improvement in noisy environments," in *Proceedings of IEEE ICASSp-2006*, Toulouse, France, pp. 493–496.